THE UNIVERSITY *of* EDINBURGH

Gene co-expression tools applied to the developing thalamus

Xavier Oliver Duocastella

Master of Philosophy

School of Informatics

University of Edinburgh

2011

Abstract

This project contributes to the understanding of how the thalamus, a major structure in the mammalian brain, develops. This is done through the creation and application of neuroinformatics tools to biological data.

The thalamus is a paired structure located ventro-caudally in the vertebrate forebrain which constitutes the most important part of the diencephalon. Its main roles are the relay of sensory information from the body to the brain, and the modulation of this information based on feedback it receives. Internally it is subdivided into nuclei, big cell clusters that belong to separate pathways. The parcellation of the thalamus into nuclei starts embryonically, where multiple genes with combinatorially overlapping expression patterns help differentiate the fate of newly formed cells.

Based on nuclei counterstain and gene expression in situ data from the Allen Developmental Mouse Brain Atlas, this project provides insights on how the mouse thalamus is parcellated at E13.5. To do so, methods are proposed to extract cell density information from reference datasets and single gene expression information. A workflow is described to unify the data from the two methods into the same anatomical space.

A method to calculate levels of co-expression between pairs of genes is then developed, providing a way to categorise pairs of genes based on their co-expression relationship in a specific brain region: no co-expression, potential, or existent. The tools implementing the methods are then applied to the combinations of several genes involved in different developmental aspects of the thalamus at E13.5 (Gbx2, Ngn2, Olig2, Otx2, Cdh8 and EphA4), to understand where cells co-express or might co-express each pair.

The results indicate that the more restricted nature of co-expression patterns of gene pairs is more useful to understand nucleogenesis than individual gene expression patterns, which might be too wide or too specific. Finally a subdivision of the developmental thalamus at this age is proposed based on the regions in common, and these subdivisions are linked to the adult thalamus.

Acknowledgements

I would like to thank John Davey, Yijing Chen, Matt Down, Martine Manuel and Chris Conway and in general the Price, Mason, Kind and Pratt research groups for helping me when I needed it.

I am very grateful to Isabel Martín-Caballero for her support during confusing times, to Jim Bednar for his advice, feedback and key support and to Tom Pratt for maintaining his involvement in the supervision even when he didn't have to.

And of course thanks to David Willshaw and David Price, for supervising my project and making this thesis exist.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Xavier Oliver Duocastella*)

To Aggeliki, for accompanying me there;
to Pat, because she knows why;
to all the friends who have been there for me;
and to Pandi.

Contents

List of Figures

List of Tables

Motivation

This is a neuroinformatics MPhil project. It aims at leading to the understanding of the development of the mammalian brain, and more specifically the thalamus, via the use of informatics tools using the mouse as model organism.

During its development in the embryonic stage, the thalamus becomes subdivided into nuclei and it establishes connections with other brain structures. While research on how thalamic tracts are established is abundant, the mechanisms of thalamic nucleation are far from being fully understood.

Tissue patterning is the procedure by which different populations of cells arise from a single one due to the influence of a range of cues. Like any other developmental aspect, it is regulated by genes being expressed in cells and their interactions, which have been suggested to work combinatorially. The study of gene expression is therefore a useful approach to learn about parcellation.

Data gathering from the thalamus is challenging and time consuming, but fortunately there are various databases containing vast amounts of biological data being shared by different institutions. The Allen Brain Institute offers the most complete mouse data: its embryonic section contains hundreds of gene expression data throughout development ready to be used.

This project will propose procedures and implement computing tools to make sense of the available data, extract and combine useful information, and ultimately gain insights on how the thalamus is partitioned during its development.

Layout

This thesis is subdivided into four chapters where the following key points are included:

1. Background: Overview of the adult thalamus and its development. Combinatorial parcellation of it.

2. Tools created in this project to extract basic information from the data. Assessment of these.

3. Tools created in this project to combine extracted information. Application to data and result aggregation.

4. Discussion of the results. Future work. Conclusions.

Chapter 1

Background

This chapter briefly describes the thalamus as it is at the end of the developmental process, its main features and connectivity and how that comes to be in the mouse. As the brain structure that this project focuses on, it is important to understand its anatomy and how the parcellation process influences its development.

1.1 The adult thalamus

The thalamus is a paired structure located ventro-caudally in the forebrain that has symmetric parts on both sides of the body and constitutes the most important part of the diencephalon. Figure 1.1 shows the location of the thalamus in the mouse from its early appearance to the fully developed late adult.

1.1.1 Internal organisation

Neurons in the thalamus are clustered, forming nuclei, as shown in figure 1.2 for the adult mouse. Usually each nucleus and its connections comprise a single functional pathway. For example, the dorso-lateral geniculate nucleus (dLGN) belongs to the visual pathway, being connected to the retina, the cortex and the superior colliculus in mammals. Initial evidence for this form of organisation and connection was found via staining of the degeneration of cells or myelin: elimination of specific functional areas in the cortex resulted in progressive degeneration of the associated nuclei [16].

Knowledge regarding the nuclei varies considerably. The most well-studied in mammals are the ones corresponding to the sensory flow of information to the cortex: dorso-lateral geniculate nucleus (visual pathway), ventro-basal and postero-medial complexes (somatosensory) and medial geniculate complex (auditory). On the other hand there are the nuclei located in the midline and others that are involved in functions that are unknown or not as straightforward to measure.

Some thalamic nuclei do not contain uniform populations of neurons but are subdivided into clusters. Main sensory nuclei are organised in cell layers, and each of them relays and processes information corresponding mostly to a segregated pathway of the signal they receive: different sound

frequencies in the auditory complex, visual features in the visual nucleus, and tactile features in the somatosensory nuclei [7, 16].

1.1.2 Description of connections

The thalamus and the cortex are very close collaborators: it is estimated that in the mammalian brain and depending on the thalamic nucleus, between 70 and 99% of neurons project to the cortex [30]. Evolutionarily, the size of the thalamus has been found to be proportional to the size of the cortex [16].

Generally speaking, each thalamic nucleus is connected either specifically to a single cortical region or broadly to several cortical regions involved in the same pathway. Electrophysiology *in vivo* and dye injections *in vitro* revealed that for some nuclei their connections map in a topographical manner to their associated cortical area [15, 1]: thalamic cells close to one another innervate neighbouring regions in the cortex (see figure 1.3). Topographic relationships have since then been discovered in connections established by several nuclei or divisions of them [16] and have been most clearly described for nuclei involved in sensory processing.

Table 1.1 contains the list of nuclei groups and corresponding nuclei or divisions existing in mammals, standardised by ref. [16]. For each of them, and from a mouse perspective, it describes its main cortical target and whether any topography has been detected with one or more of its targets. The compiled data shows how nuclei groups or subgroups tend to target same or related cortical areas and how in some cases the main nucleus in the group innervates the main associated cortical region and the remaining ones connect with secondary or surrounding regions.

Figure 1.1: Location of the thalamus in the mouse brain on saggital sections at ages E13.5, E16 and P56 from top to bottom.
In the top figure the precursor regions that will give rise to thalamus and prethalamus have been labeled as Th and pTh. During development the rostro-caudal axis, used for coronal sectioning, becomes more similar to the antero-posterior axis, defined by the longitudinal neural tube (see fig. 1.8). By adapting this sectioning angle to each age it is possible to relate the adult position of thalamic nuclei or divisions to the embryonic position.
Top figure adapted from the Allen Developing Mouse Brain Atlas (`http://www.brain-map.org` . For the bottom two figures, copyright acknowledged to the Gene Expression Nervous System Atlas (`http://www.gensat.org`).

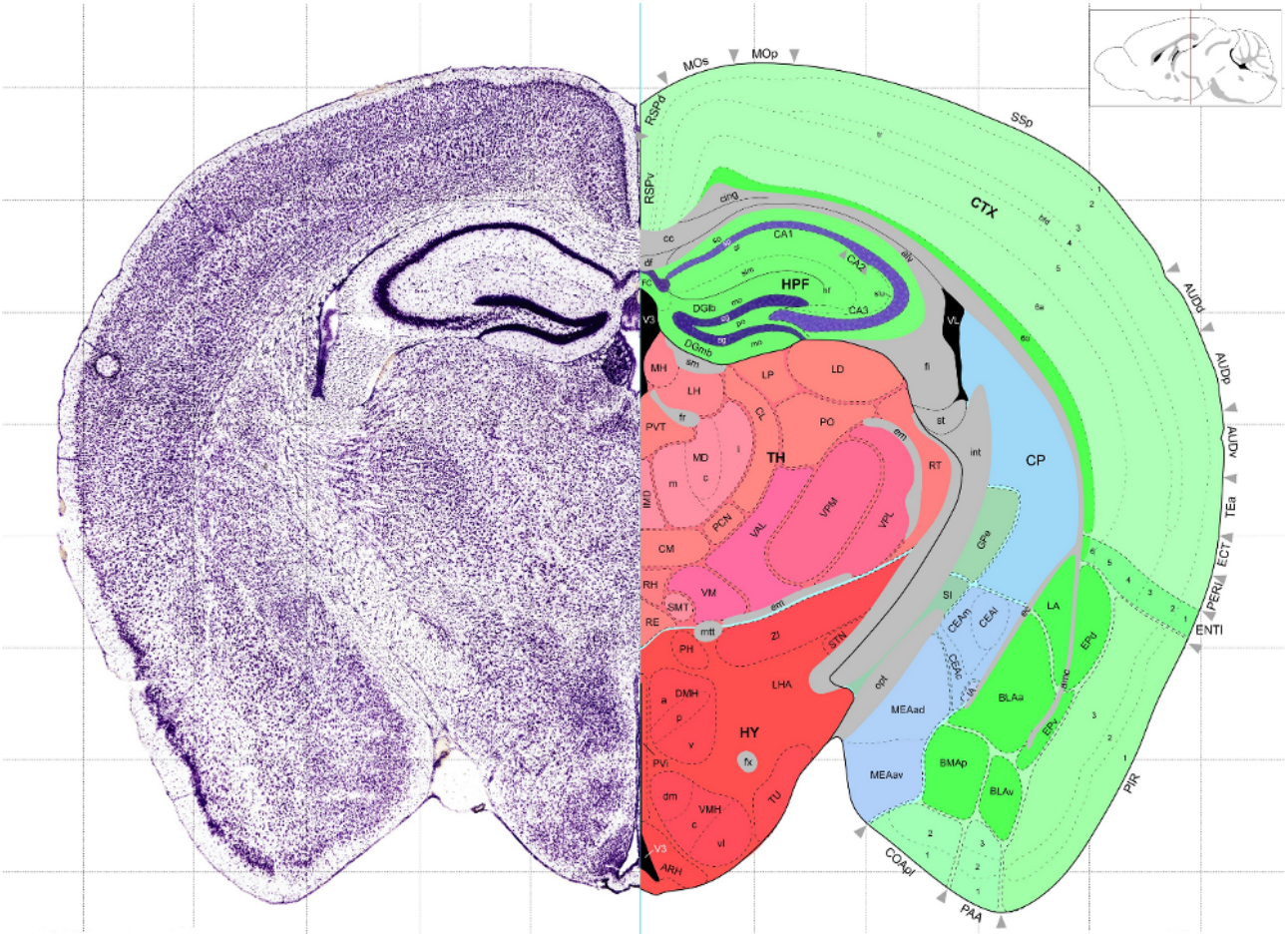


Figure 1.2: Internal layout of the thalamus in an annotated adult slice.
The different shades of pink on the right half correspond to different nuclei groups from thalamus and prethalamus. Note that this view also includes part of the cortex (green), hypothalamus (red) and cortical plate (blue). Copyright acknowledged to the Allen Mouse Brain Atlas (`http://www.brain-map.org`).

1.1.3 Function

Being on the pathway of sensory data on its route to the cortex, the thalamus has been classically described as a relay of information.

Research in the last decade has linked the thalamus to an active role regarding its interaction with the cortex and even regulating cortico-cortical communication. It is now accepted that the thalamus filters information coming to the cortex based on cortical feedback with the involvement of the reticular nucleus [29]. Intralaminar and midline nuclei have for example been linked to attention, arousal and awareness [8], and clinical research using imaging techniques has strongly helped identify thalamic influence in a variety of diseases. Furthermore, modeling efforts to functionally study the thalamus given the properties of its cells and interaction with other diencephalic structures [30] are also creating debate and giving life to research in the area, as well as contributing to the general understanding of how the thalamus works by uncovering the general principles behind it.

Cell types

Neurons in the thalamus fall into two categories: relay neurons and interneurons.

Relay neurons are glutamatergic and produce excitatory actions that drive their targets closer to firing. They transmit information to higher targets from different pathways via their afferents, with the cortex as the most important destination. Interneurons are on the other hand GABAergic cells with a short range of inhibitory activity that receive input mostly from the cortex and the thalamus itself and have been found to modulate the signals from the relay neurons [30].

1.2 The developing thalamus

Origin

The thalamus is the main structure in the diencephalon, which is formed from the hypothalamus, the prethalamus and the epithalamus.

Being part of the central nervous system, the diencephalon develops from the neural plate. Both the diencephalon and the telencephalon form the forebrain, one of the three primary vesicles in the neural tube along with the hindbrain and the midbrain [12, 5].

1.2.1 Tissue patterning

During development, newly generated cells migrate to their destination, associate with similar cells and become highly specialised. Throughout this process they are exposed to and produce multiple extracellular signals that in turn alter their behaviour.

Nuclei group	Nucleus/division	Main cortical target	Topography
Ventral	VPM	S1 mostly, some S2	Y
	VPL	S1 mostly, some S2	Y
	VPi*	?	-
	Basal VM	Taste, body sensors	?
	VLc	M1	Y
	VM	?	N
	VA	?	N
Medial geniculate complex	vMGc	A1	Y
	dMGc	A1 surroundings	N
	mMGc	A1 and surroundings	N
Dorsal lateral geniculate	LGN	V1	Y
Lateral posterior and pulvinar	LP	Area 17, around visual area	N
	Pu*	Temporal CX, extrastriate visual	-
Posterior group	POm	S1 (anterior), S2 (posterior)	Y
Intralaminar (anterior group)	CM	Medial+basal CX	?
	PC	Lateral CX	?
	CL	Lateral CX, sensorimotor+	?
	Rh	?	?
Intralaminar (posterior group)	PF	Rhinal sulcus, cingulate gyrus	?
	CeM*	M1	-
Medial complex	MD	Olfactory,frontal lobe	?
	PT	Olfactory,frontal lobe	N
	MV	Entorhinal CX, hippocampus	?
Anterior+lateral dorsal	AD	Retrospenial granular, subicular areas	?
	LD	Areas 18b and 19c	?

Table 1.1: Survey of thalamic nuclei groups in mouse.
For each individual nucleus, information regarding its main afferent target, and whether any kind of topography is evident. *Asterisk*: nucleus cannot be identified in mice. *Question mark*: data not known (third column) or no evidence found (fourth). Compiled from ref. [16] with additional references cited in ref. [16].

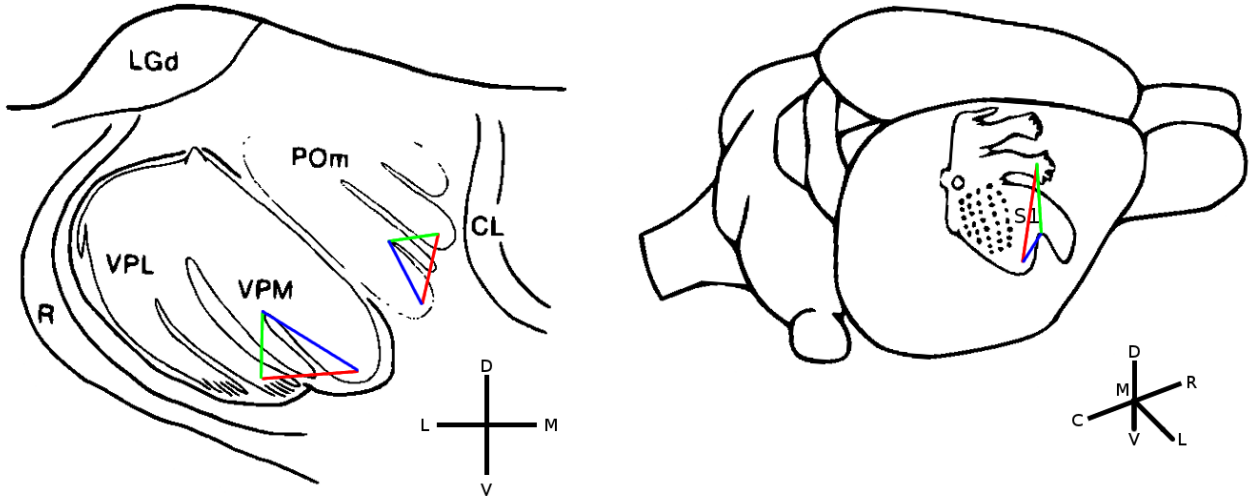


Figure 1.3: Topographic relationship between main somatosensory nuclei in the thalamus and cortex in the rat brain.
Points on the VPM, VPL and POm in the coronal thalamic slice (left) map continuously to S1 in the cortex, shown on the whole-mount brain (right). The overlaid triangles show, for the specific subregions they cover, that the spatial relationships between the edges and vertices are maintained in the different projections.
Thalamic nuclei: LGd dorso-lateral geniculate, POm postero-medial, VPM and VPL ventro-posterior lateral and medial, R reticulate (prethalamus), CL central lateral. S1: primary somatosensory cortex. Adapted from ref. [10].

Morphogens

A crucial type of signal for development, the morphogens, is a class of proteins that induce the expression of sets of genes in cells, effectively determining their fate.

Morphogens are expressed in signaling centres, transient structures in key places from where development of the area is induced. They diffuse as a gradient, and affect target cells in different ways depending on the level of exposure to them (fig.1.4). This makes morphogens key components for starting the cascade of events that results in differentiation and patterning of tissue areas.

Transcription factors

Proteins that bind to DNA and affect its transcription into mRNA by either promoting it or repressing it are called transcription factors, or gene regulatory proteins.

Transcription factors are extremely useful when a cell is required to alter its behaviour. Some can enable the expression of a set of genes, starting a whole cell program by switching some of them on and off in a coordinated fashion. It is also possible to have a transcription factor triggering a chain of downstream genes, a mechanism that is very common when cells specialise during development.

The history, location or neighbours of a cell affect how it reacts to a given transcription factor. This variation in behaviour means that over time great cell complexity can be reached by the close collaboration of several transcription factors: the effect of having been exposed to different combina-

tions of them over time means resulting cell populations will develop differently. Figure 1.5 provides a visual example of how a limited number of these genes can contribute to many kinds of cells.

1.2.2 Patterning of the diencephalon

The development of similar cell fates for cell populations based on exposure to signaling centres, or patterning, involves genes from the homeobox family. These have been observed patterning the *Drosophila Melanogaster* embryo and determining the exact anatomical region that developed in the area where they were expressed (fig.1.6). Mammal homolog genes also have a great role in development, but in a more subtle way.

A variety of species

Research on the development of the vertebrate diencephalon studies different factors with different animal species, usually chick, zebrafish, mouse and rat. The available literature, and specially reviews ([19, 27]), combines this varied evidence and uses it to build a model of development, without acknowledging inter-species comparison problem.

The following account of diencephalic and thalamic development will therefore use examples from different species to provide a unified view, similarly to the literature.

Axes

Neural tube derivatives in the hindbrain develop local regionalisation based on combinations of restricted domains of signaling genes expressed longitudinally (Antero-Posterior axis) and tangentially (Dorsal-Ventral axis).

This method of spatial definition of boundaries or fate has been observed in the diencephalon too, so the description via the two axes is the best approach to describing how the diencephalon is subdivided into the different thalami and how these might be divided too [19].

Dorso-ventral axis

When the neural tube is formed by the folding neural plate, most of its cells are neural precursors. The only exception to that are two groups of cells at each dorsal and ventral end of the tube, which differentiate into two strips of cells: the roof and the floor plate respectively (see fig.1.7).

Similarly to the spinal cord, the roof and floor plates are the key transient signaling areas for the DV patterning in the prospective diencephalon, and as a result of their signaling influence, they help define another two divisions in this axis, the alar and basal plates.

Cells in the floor plate secrete a range of signaling molecules that have been linked to development in other areas of the nervous system: Shh, Fgfs, Bmps and Egf-cfcs as well as netrin. Cells in the ventral midline area of the roof plate express fewer signaling molecules, mainly Bmps and Wnts. While

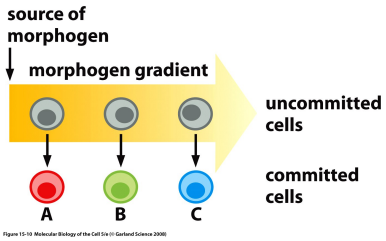


Figure 1.4: Effect of the exposure to various levels of morphogen expression.
Here different exposures induce different gene expression and fate. Copyright acknowledged to ref. [2].

the main three genes expressed in the floor plate are consistently required for correct development of the diencephalon, inhibition of the signaling molecules in the roof plate affects the development only partially [19].

Finally, cells from the basal and alar plate are not involved in signaling activities and will form most of the thalamic nuclei [19].

Antero-posterior axis

The diencephalon is bounded anteriorly and posteriorly by the telencephalic-diencephalic and diencephalic-mesencephalic boundaries respectively (DTB and DMN).

This axis also contains the zona limitans intrathalamica (ZLI), an important signaling centre situated between thalamus and prethalamus (see fig.1.8). Recent evidence has shown that the ZLI acts as a local organiser for the region and has been called mid-diencephalic organizer (MDO) [27].

Being a local organiser, it secretes signaling molecules, keeps cell populations separated, is required for proper development of the region and if it is transplanted ectopically it induces the same fate to the neighbouring region as it does in its original location. Figure 1.9 illustrates the process of MDO formation as well as how it helps shape the thalami in the zebrafish.

Induction of the MDO starts at E8.5 in the mouse. Prethalamic and thalamic anlages express Fez and Otx respectively (fig.1.9.A) and are believed to induce Shh expression in the sharp border between them [27] (1.9.B). Maturation of the MDO happens at E10.5, when it fully separates the prethalamic and thalamic and it is also surrounded by expression of Nkx2.2[18]. The thalamic area also expresses Irx, and together with Fez on the other side they maintain the MDO borders via repression [26] (1.9.C).

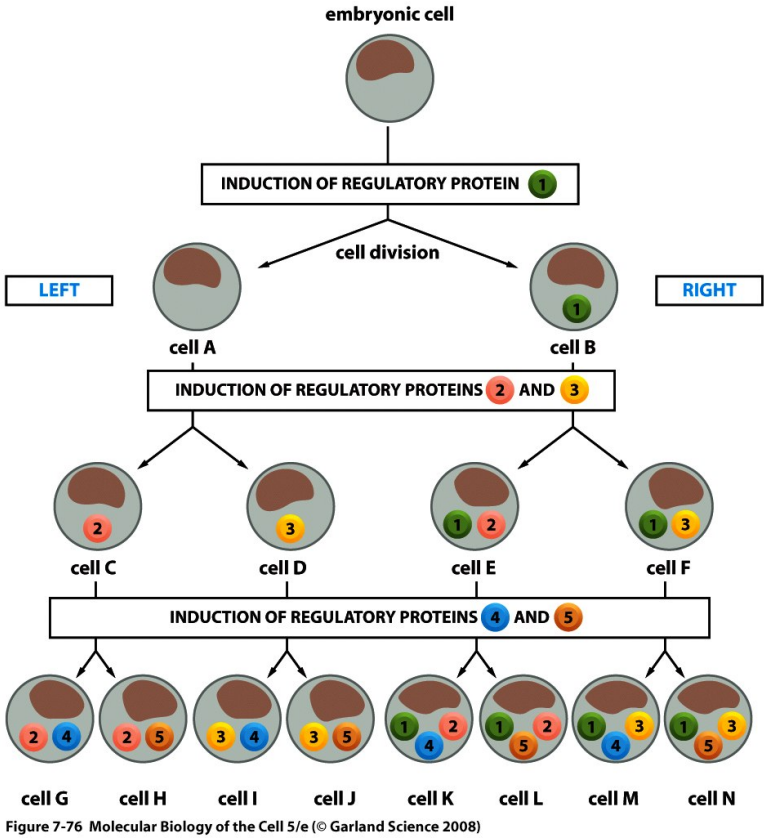


Figure 1.5: Cell fate specialisation via combinatorial expression of transcription factors (gene regulatory proteins).
In a very simplistic way, during each cycle the cells react differently to the induction signals depending on their different location relative to their parent (left/right). Fates are not decided by a single regulatory protein, but a combination of them. Copyright acknowledged to ref. [2].

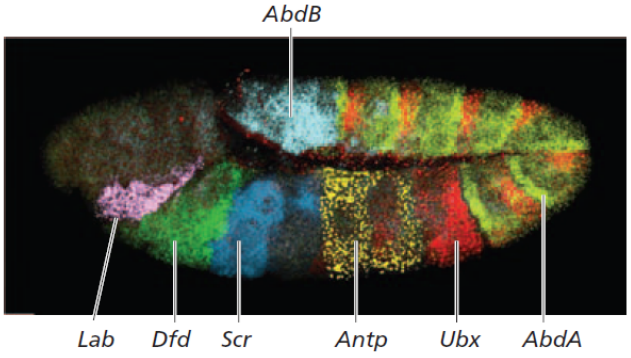


Figure 1.6: Expression of homeobox genes in the embryonic *Drosophila Melanogaster*.
Pseudo-coloured areas, corresponding to the expression pattern of each gene, will give rise to a specific part of the adult fly, such as antennae, thoracic sections or pairs of legs. Copyright acknowledged to ref. [2].

1.2.3 Parcellation of the thalamus

When the MDO is established, it expresses Sonic Hedgehog (Shh) and this causes cell differentiation in both thalami.

Progenitor stage During its early stages of development (E10.5-E12.5), the thalamus can be divided in two progenitor domains located caudally from the ZLI: pTh-rostral, which expresses Nkx2.2, Mash1 and Olig3; and the pTh-caudal, which expresses Neurog1,Neurog2 and Olig3 [38] (fig.1.10).

The transcription factor Her6 acts as a switch for fate selection in zebrafish for these domains (fig.1.9.D). Her6 is expressed in the prethalamus and on the far rostral part of the thalamus, and causes cells exposed to Shh to express Mash1 (requires Her6) or Neurog1 (repressed by Her6) [25]. Expression of these genes leads in turn to the differentiation of GABAergic -inhibitory/modulatory- neurons (pTH-R) and glutamatergic -relay- neurons (pTH-C) respectively [38, 25].

Morphogens have been found that alter the balance between these two domains: thalamic nuclei become shifted in the rostro-caudal axis after Shh or Fgf8 are either knocked out or over-expressed [39, 17]. Suppression of Otx2 for example prevents cells from becoming GABAergic, activates other genes that are found in the pretectum (Pax3,Pax7,Lim1), and increases proliferating activity [23].

Postmitotic stage Cells that enter the post-mitotic stage migrate laterally out of the ventricular zone located in the midline towards their destination within the growing thalamus.

Some thalamic nuclei can be identified as early as E14.5 [3], but it is not until early postnatal age that all can be distinguished clearly. A study done in 2001 looked at the expression patterns of some LIM-homeodomain transcription factors (Isl1, Lhx1, Lhx2, Lhx5 and Lhx9) and genes that are important for development (Gbx2, Ngn2 and Pax6) during late development (E10.5-E16.5 and P2) and compared them during the different stages. The findings were that, save obvious anatomical variations due to growth, their expression domains were stable during the ages studied and so did the relationships they kept. Furthermore, their multiple overlaps within the thalamus were found to mark prospective nuclei or groups of nuclei at P2, suggesting that these genes acted in a combinatorial fashion to help parcellate the thalamus [21]. Table 1.2 summarises the early postnatal findings of the study.

Only more recently, two related studies took a similar approach and studied expression patterns of 32 genes at E10.5-E12.5 and compared them with E15.5 and P6. These included BMPs, guidance cue and cell adhesion genes, homeodomains used in the previous paper and other genes identified as potentially useful. The result was a classification of most of the genes according to where and when they were expressed in the developing diencephalon or the early thalamic subdivisions, and as a consequence more candidate genes were proposed to justify nucleogenesis [34, 40].

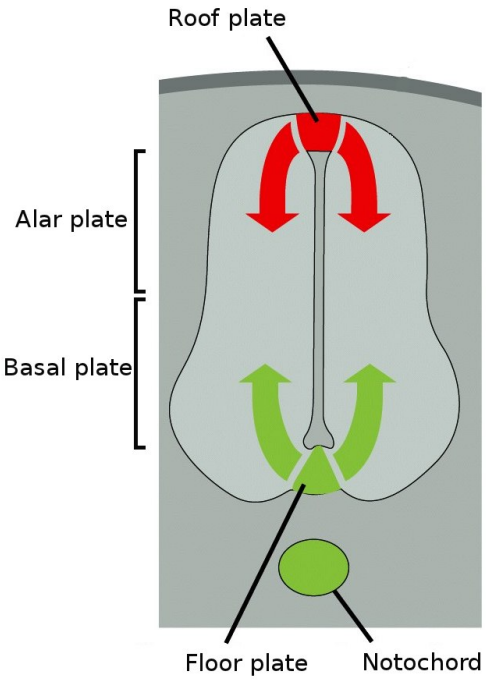


Figure 1.7: Dorso-ventral patterning in the central early diencephalon of the mouse.
Red: expression of TFGs, BMPs and retinoic acid. Green: expression of Shh, Fgfs and retinoic acid. Copyright acknowledged to ref. [2].

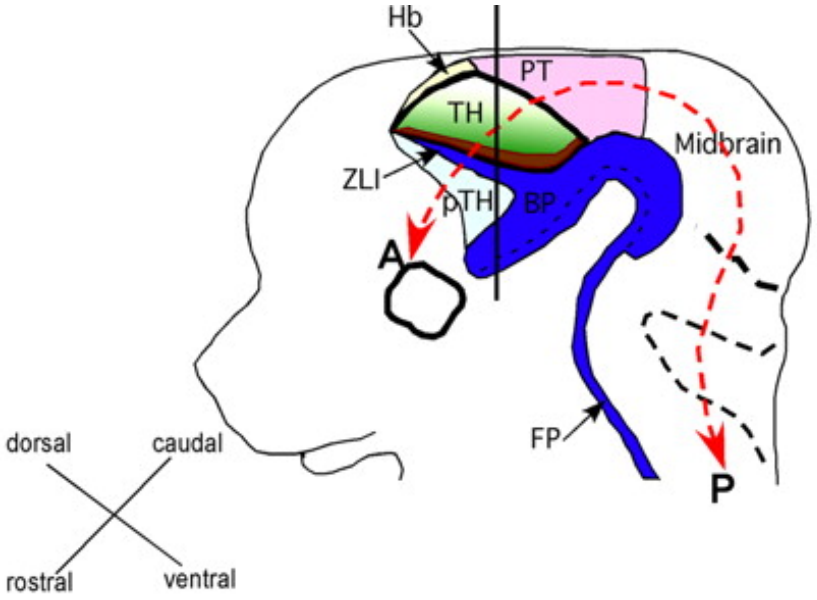


Figure 1.8: Thalami layout in the E12.5 mouse according to the Antero-Posterior axis defined by the longitudinal neural tube.
Note that the rostro-caudal axis does not match the antero-posterior one throughout the whole span of the neural tube. The rostro-caudal and in this case dorso-ventral axes refer to the sectioning of the tissue.
TH: thalamus, pTH: prethalamus, Hb: habenula, ZLI: zona limitans intrathalamica, BP and FP: (neural tube's) basal and floor plates, PT: pretectum. Copyright acknowledged to ref. [39].

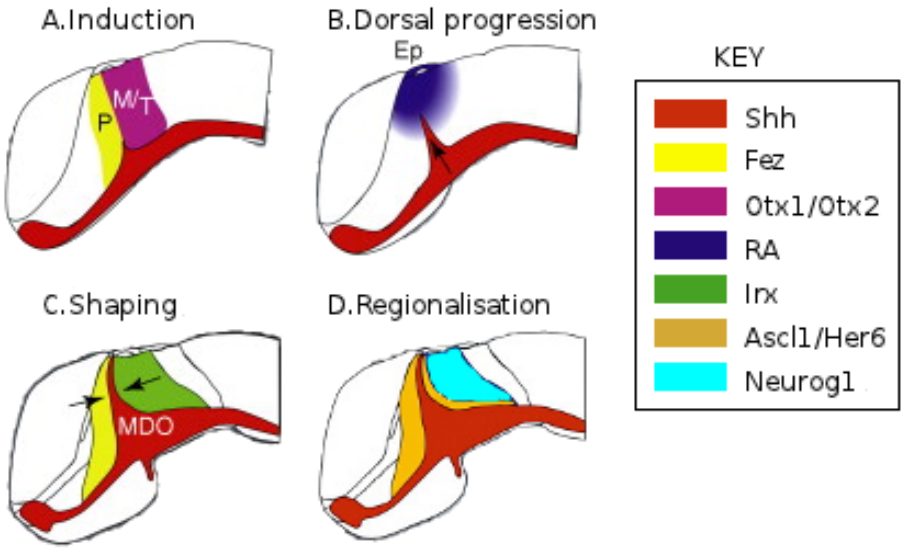


Figure 1.9: Development of the mid-diencephalic organizer (MDO) in the zebrafish and its role in thalamic development.
Fez and Otx expressed in thalamic and prethalamic anlages induce Shh expression (A), which progresses dorsally across the border until a repressing agent (RA) blocks it (B). The MDO is shaped via expression of Fez and Irx on each side (C) and when it is mature its Shh causes different cell fate by activating either Ascl1 via Her6 or Neurog1 (D). P: prethalamic anlage, M/T: MDO/thalamic anlage, Ep: epithalamus. Adapted from ref. [27].

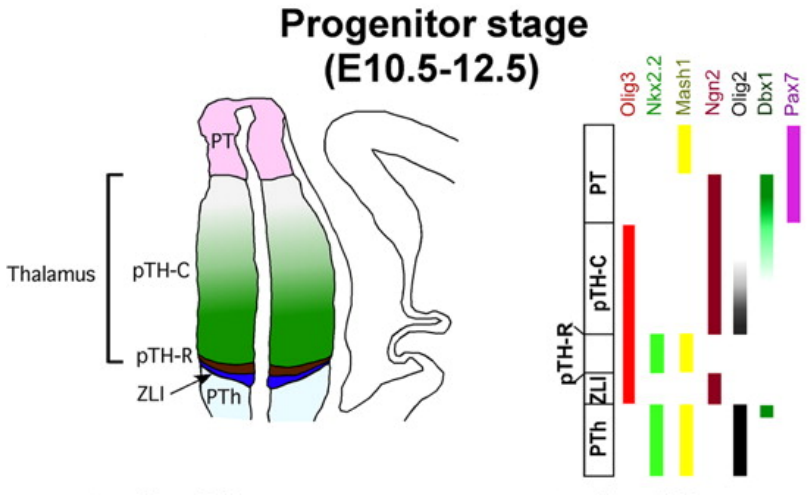


Figure 1.10: Thalamic progenitor domains in the mouse as defined by [38].
Left: coronal slice of the brain in figure 1.8 showing the different domains present during the progenitor stage. Right: expression of different marker genes in those domains as found by ref. [38] and expanded in ref. [39]. pTH: thalamic progenitor domain (rostral or caudal); PT: pretectum; ZLI: zona limitans intrathalamica (MDO); PTh: prethalamus. Copyright acknowledged to ref. [39].

	Main sensory			Other nuclei												
	VP	LGN	MGv	AM	AV	CL	CM	LD	MD	MGd	LP	Po	PP	PT	PV	VL	VM
Lhx2	-	-	+	*	-	+	+	-	-	*	-	-	+	-	+	+	-
Lhx9	*	+	*	+	+	+	+	+	+	+	+	+	+	+	+	+	-
Gbx2	-	-	+	*	-	+	+	-	+	+	+	-	-	+	+	-	-
Ngn2	*	*	*	+	+	-	*	+	-	-	-	-	-	-	-	*	+
Unique	Y	Y	Y	Y	N-1	N-2	Y	N-1	N-3	Y	N-3	Y	Y	N-3	N-2	Y	Y

Table 1.2: Differential expression of LIM-homeodomain and other transcription factors in P2 mouse thalamus.
For every gene its level of expression in each of the prospective nuclei areas is indicated. *Plus sign* (+): expressed; *asterisk* (*): weakly expressed; *minus sign* (-): not detectable.
The bottom row specifies whether the nucleus has a unique combination of gene expressions (Yes/No) in each prospective nucleus area. If genes share a combination an arbitrary number to link them is specified. Adapted from ref. [21].

Chapter 2

Data extraction

To study how gene co-expression shapes the developing thalamus it is necessary to first measure individual gene expression. A confound does exist when studying gene expression and it involves cell density: given two tissue selections with high and low cell density levels where all the cells express a given gene, it would seem that the level of expression is higher on the densest area when it is actually the same. To address this confound cell density will also be measured.

Algorithms are proposed to extract information from the data in this chapter, and tools incorporating them are implemented so that the large data available can be processed in a consistent way.

2.1 Cell density information

The calculation of cell density will be based on cell identification in the source images. A method to obtain this data will be discussed, implemented, and compared with currently used density assessments.

2.1.1 A morphological method to measure cell density

The proposed procedure is based on morphological features and relies entirely on the calculation of the *Hessian* matrix of the image, a square matrix of second order partial derivatives of a function. Derivatives provide useful information about function morphology, such as values for steepness (first order) and curvature (second order) at specific points. When applied to images, the calculated values become the means to locate specific geometrical features [24]. In the specific case of the Allen Developmental Brain Atlas (ADBA), cells have borders that are clear enough to suggest that morphological information can be extracted successfully.

The idea for an approach based on the Hessian matrix originated from previous work done on my Masters [22] and was initially supported by a paper on biomedical image processing [14]. In it, the author defined a lung cancer nodule detection algorithm in computed tomography scans. The objective was to develop a computer-aided detection method to assist imaging technicians in early

lung cancer detection. To do so, the resulting algorithm applied a combination of Hessian-based processing with a variable pre-processing scale range to detect a whole range of sizes in the data.

Procedure

Images in the ADBA have their information stored in RGB, three integer values in the range [0,255] corresponding to Red, Green and Blue colour channels.

To enable derivative calculations floating point values are needed, so an initial pre-processing step is to convert the initial image to grayscale floating point values.

Hessian matrix. The pre-processed image can be expressed as a real-valued two-dimensional function $f(x,y)$ where the output value range is constrained by the maximum range of the imaging equipment. The resulting Hessian matrix for this two-dimensional input is:

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

This symmetric matrix relates the evolution of values in one axis to the others including itself. Centered at each input data point it provides information on how the value of that point changes regarding the neighbours in the space around it.

Eigenvalues. The next step is to calculate the eigenvalues of the Hessian matrix. It is done by resolving the following equation for each point, where I is the identity matrix, λ corresponds to the vector containing the three eigenvalues (since data is three-dimensional) and det is the determinant of the resulting matrix:

$$det(H - I\lambda) = 0$$

The set of values (λ_1, λ_2) for each input data point are then ordered according to their value. This results in two datasets, corresponding to large and small eigenvalues, which have different influences and point out different features of the data.

Classification. The final step of eigenvalue calculations is to apply a classifier; a function that relates eigenvalues for any given point and returns a number that allows sorting of points according to whether specific higher-level features are represented.

Classifiers are defined empirically by exploring the data. In the case of ADBA counterstained tissue, the classifier devised works by accepting points where both eigenvalues have values over 0 and zeroing the rest. Since by definition $\lambda_2 \geq \lambda_1$, only the smallest eigenvalue is required. The value C of the classifier for a given point is

$$C = \begin{cases} 0 & when\, \lambda_1 \geq 0 \\ 1 & when\, \lambda_1 < 0 \end{cases}$$

The resulting classifier value for each point of the image identifies those belonging inside two-dimensional clusters with round or oval shape.

Filtering. These clusters correspond usually to cells, but there are two situations that justify an extra filtering step to ensure the quality of data.

The first case is white gaps. Areas of no tissue surrounded by cells are also accepted by the algorithm since they satisfy the same morphological conditions as cells. To keep these cases out, the points need to be compared to their associated value in the initial tissue image. If the points of the cluster have brightness greater than any counterstained tissue they are discarded.

The second identifies spatial noise. Clusters of all sizes are usually detected, including some that do not represent whole cells or are simply imaging artifacts. Because detected real cells have many pixels classified as valid, false positives can be ruled out by removing those that have only one or two valid pixels.

Neighbouring pixels corresponding to the same cell are then shrunk so that only a single point located in the middle of the identified cluster remains. This makes any future calculation such as method comparison much easier.

Assessment

Initial tests where the proposed cell detection method was applied to the available data were encouraging. These were done in local thalamic regions that were useful when fine tuning the procedure, but which are far from the size and variability of the data that will be ultimately used.

It is not simply possible to assess the quality of any cell detection procedure properly. There are no means of obtaining *true* values to validate against for a given data set, because every approach available grasps the underlying truth in different ways. Therefore, it is possible to compare results obtained from different methods to assess how similar they are. This, in turn, provides an idea on how closely they might be representing the existing data.

The Hessian-based method will be tested against two approaches to detect cell bodies and gather cell density in their respective scales: small and big.

2.1.2 Small scale assessment: manual annotation

To study how well the method works in the small scale its results are compared with data generated by manually annotating representative data. Evaluation criteria are defined and then applied to the automatic results obtained by processing the same data.

A focus of interest is understanding what the best source of information from the image is. Despite having a yellow (red+green) counterstain, each channel contains different information and using the right one might yield more accurate results.

Data used

The tissue sections used belong to the reference coronal sets of the ADBA. These are $20\mu m$ thick and are stained with Feulgen-HP Yellow DNA counterstain, which reveals cell nuclei. 100x100 pixel areas of the thalamus are used. Due to the calibration used in the imaging process, a pixel is equivalent to $1\mu m^2$, so images are effectively $10000\mu m^2$ in size.

Three data sets were used, corresponding to embryonic ages E13.5, E15.5 and E17.5. E18,5, a more common age to study development than E17.5, could not be used because its counterstain was different. Each set contained three images representing one of three density categories : low, medium and high. These density categories were defined visually and correspond, as seen *a posteriori* (table *2.1*), to approximately 33%, 75% and almost 100% of all the image pixels containing counterstained tissue. Counterstained nuclei sizes range from 50 to $75\mu m^2$, which implies a nucleus radius between 4 and $5\mu m$ under the assumption that they have circular shape.

Also, the pixel value used to threshold images and separate pixels containing tissue information from background is nearly constant in all images. This is due to the Allen Institute using automated protocols to image experimental data, and it is very useful to in turn allow automatic analysis of that data.

Manual annotation

Manual annotation of the data sets is done with the original colour image using Fiji (ImageJ distribution). Cells are marked by placing a single black dot in the center of the counterstained nuclei that are detected via multiple zooming in and out of the image. All the images are processed during the same session to ensure consistency of criteria. Figure 2.1 shows the results of manually annotating the first set.

Even though the resolution of the data is remarkably high, there is still a lot of noise and blur on the potential structures. A conservative approach was employed and only areas that looked clearly valid were annotated, for example those with a distinguishable border or shape. Some cells were probably missed, specially in the pictures corresponding to high density, illustrating one of the shortcomings of this method.

Table 2.1 contains the results of the manual annotation procedure, all generated after the annotation was finished. Images in each density category have point counts that are not too far apart and, as commented in the previous point, have similar percentages of pixels belonging to tissue. The estimated point count based on an average circular nucleus radius of $4.5\mu m$ and the percentage of tissue pixels is also reasonably close to the actual number of points detected.

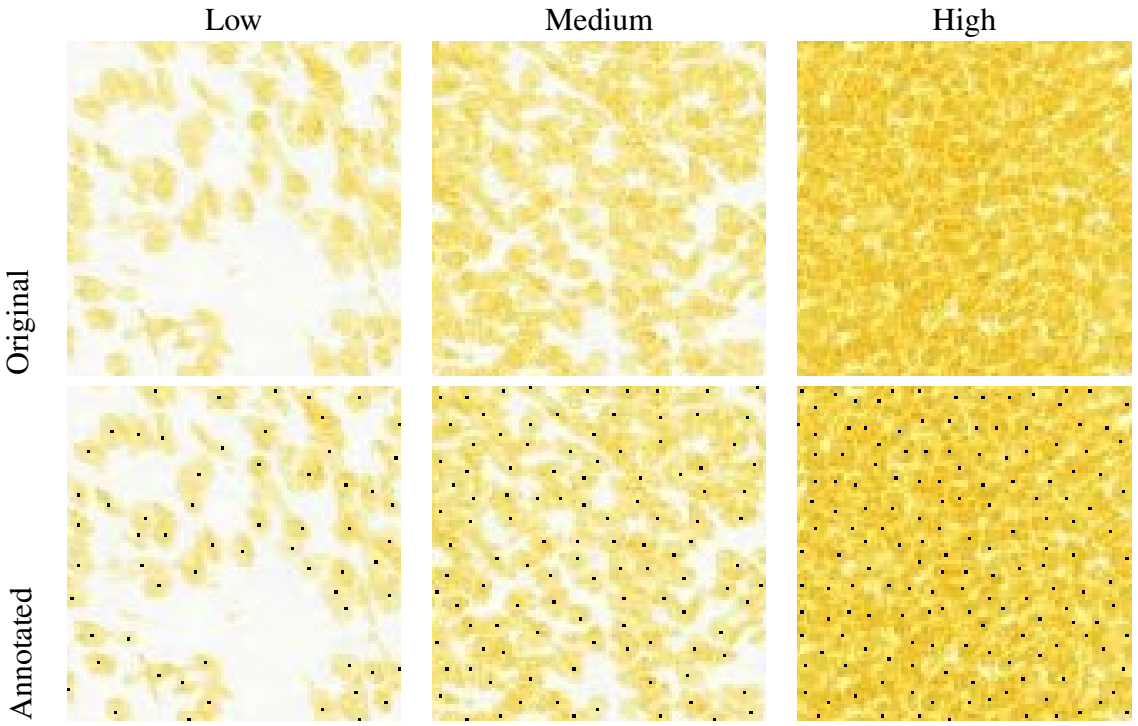


Figure 2.1: Manual annotation of datasets.
Data set #1, corresponding to different E13.5 thalamic sections. Manual annotation of the data set for the three density values. Zoom: 125%.

	Set 1 (E13.5)			Set 2 (E15.5)			Set 3 (E17.5)		
	L	M	H	L	M	H	L	M	H
% of pixels are tissue	33	72	100	38	74	100	29	80	100
Estimated number for 4.5μm radius	53	113	157	60	116	157	46	126	156
Manually counted points	63	112	161	67	131	155	62	134	183

Table 2.1: Details regarding data sets and their manual annotations.
Rows contain, in order, the percentage of the image corresponding to stained tissue, the estimated number of cells present given the previous ratio and an average nucleus with radius of 4.5μm (surface of 63.62μm^2), and the number of manually annotated cells.

Criteria for comparison

To understand how well the automatic method works its results will be compared with their manually annotated counterparts. Two criteria are used to that end: nearest neighbour distance and hit rate.

Nearest neighbour distance. The distance between a cell nucleus identified manually and the nearest neighbour of those detected automatically is used as a measure of how likely it is that they correspond to the same feature.

At the data set level, the average distance from each point of the first data set to its nearest neighbour in the second is a useful indication of similarity: the lower it is, the better the points in the sets

match. As a consequence of the differing approaches taken by both annotation methods, two data sets A and B whose points match accurately will still probably have an average distance to nearest neighbour higher than zero. Data set B could be seen as A with some noise added to each of its points or viceversa.

In this case, automatically picked points are compared to manual ones, defined by putting a dot in the center of each nucleus. Because the manual annotation procedure does not yield exactly the same results every time it is re-applied, variations are to be expected in the average nearest neighbour distance when comparing datasets. It is therefore important to see what the influence of these small changes in manual annotation would have on the distance as a way to determine the level of similarity between the two groups.

Figure 2.2 shows the variation of the average distance between each manual data set and a version of itself in which points have been shifted randomly a number of pixels inside the range [-DisplacementFactor,DisplacementFactor] in both X and Y axes. Each manual data set has been used, and the factors range from zero to nine pixels.

When increasing the factor the distance saturates more with higher density. Having a high number of points means that at some point their random displacement will move them to other points' position, reducing the nearest-neighbour distance despite having been moved further.

Nuclei radii have been empirically measured to be between 4 and 5μm. By using this value as the largest displacement factor allowed, a maximum average nearest neighbour distance of ~3μm indicates that the second dataset is within radius distance and therefore they are both very similar.

Hit rate. The nearest neighbour distance criterion could accept a bad dataset: the method might find many automatic points close to a manual one, giving a low distance measure but leaving others unmatched (shown in figure 2.3).

A solution is not only to examine how near points from the two groups are, but how many of the points in the manually annotated dataset have been successfully identified by the automatic method. The hit rate is determined by applying the nearest neighbour distance calculation described before for each point. A point is then classified as *hit* if its nearest neighbour measure is lower than a nucleus radius.

Calculating the hit rate is therefore a useful way to measure quality of data matching, and is used to provide more insight to the nearest neighbour figures.

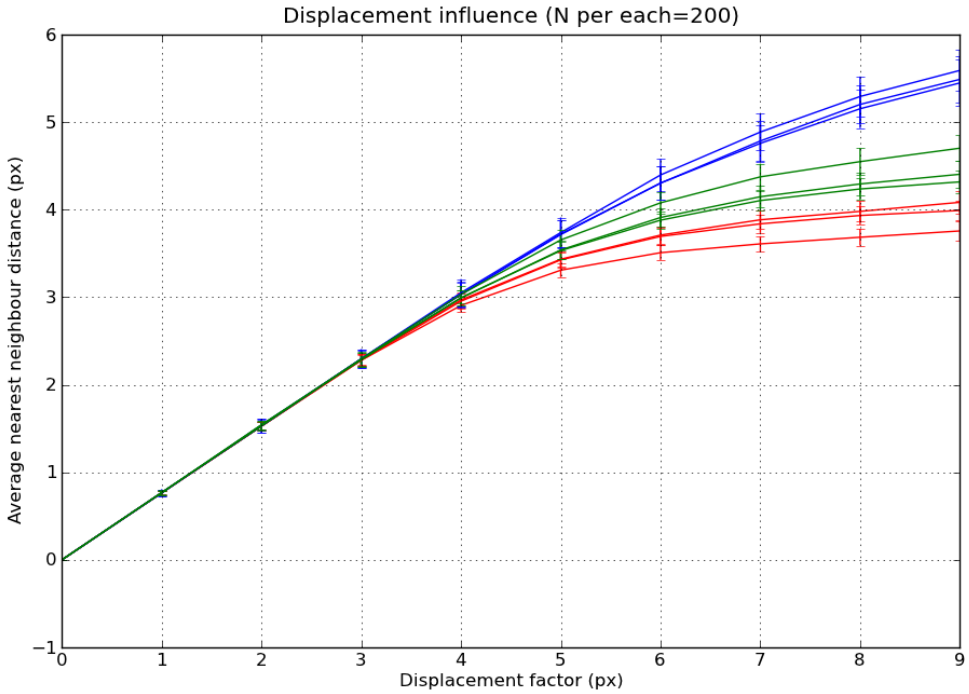


Figure 2.2: Effect of random displacement of the manually annotated points [X axis] on the mean nearest neighbour distance [Y axis].
The Red, Green and Blue lines correspond to the three high, medium and low density datasets respectively.

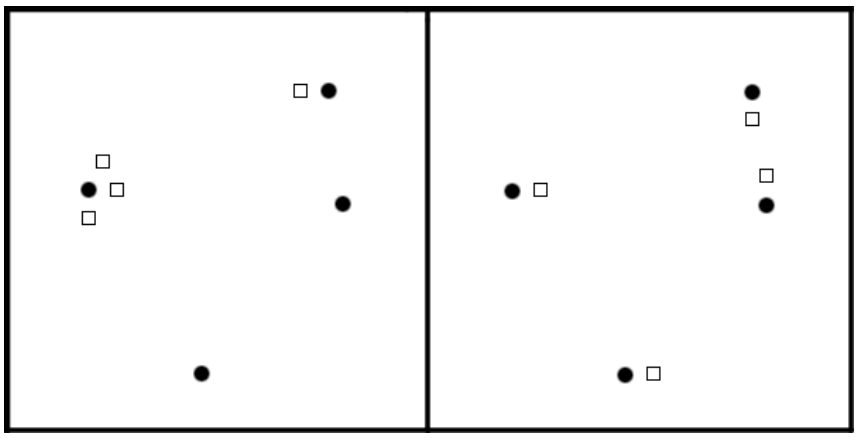


Figure 2.3: Hit rate as a complement of average nearest neighbour distance.
Sample situations where the black dots belong to the manual dataset and the squares to the automatic dataset. In both cases the average nearest neighbour distances from automatic to manual points are similar, but the hit rates are 50% on the left and 100% on the right.

Automatic *versus* manual annotation

The protocol for the analysis starts by separating the information of the image used, the colour channel (Red, Green, Blue, Gray), for each of the test images. The Hessian-based procedure is applied to them to detect points. These and the corresponding manually annotated points are processed with the comparison software to study the criteria previously described. Figure 2.4 shows the automatic points obtained in the four information sources of a data set.

As described previously, for each pair of annotated images there are two groups of results: the average nearest neighbour distances and the hit rates. The calculations are used in both directions, so the relationships (manual->automatic) and (automatic->manual) for distance and hit rate are considered. This helps compensate for channels that yield huge amounts of automatic points: it would appear that their data matches the manual points more accurately, but the excess number of points might make other moderate channels preferable.

Tables 2.2 contain the summarized results according to the density category of the datasets (A) and the channel used as source of information (B). Table 2.2.A serves as an indication of how the Hessian-based method copes with data corresponding to different cell densities. Both for average distance and hit percentage in the $m \Rightarrow a$ case the quality of the outcome is high for low density but it degenerates for higher values, while the $a \Rightarrow m$ case is exactly in the opposite situation. This could be due to the automatic method finding too many points with lower density, and not enough with higher. Medium density, previously linked at around 75% of the pixels containing tissue, has the most similar results, suggesting its results are the most reliable ones.

Table 2.2.B is helpful to see which source of data is the most helpful to obtain the best quality of information. In addition to average nearest neighbour distances and hit rates an additional indicator is used: “Wins”, which measures the number of datasets that the specific channel has offered the best hit rate for. This takes into account a margin of variation of 4% to allow for close competitors to be considered good options too.

The Red channel offers the best numbers of all in the $m \Rightarrow a$ case because it is the source of data that generates most points. This is visible underneath, where the distance is the highest and the hit rate and number of Wins are the lowest. The Blue channel has the opposite situation. Since it provides fewer points, the $m \Rightarrow a$ case is worst than the $a \Rightarrow m$ one, and the averages of the two are still not that good.

The Green and Gray channels are very similar in all aspects, and do indeed provide the best and most balanced results. Given that both channels are equally useful, and since it is easier to write software tools that interact with grayscale versions of images, this will be the chosen channel to convey information to be processed with the Hessian-based procedure.

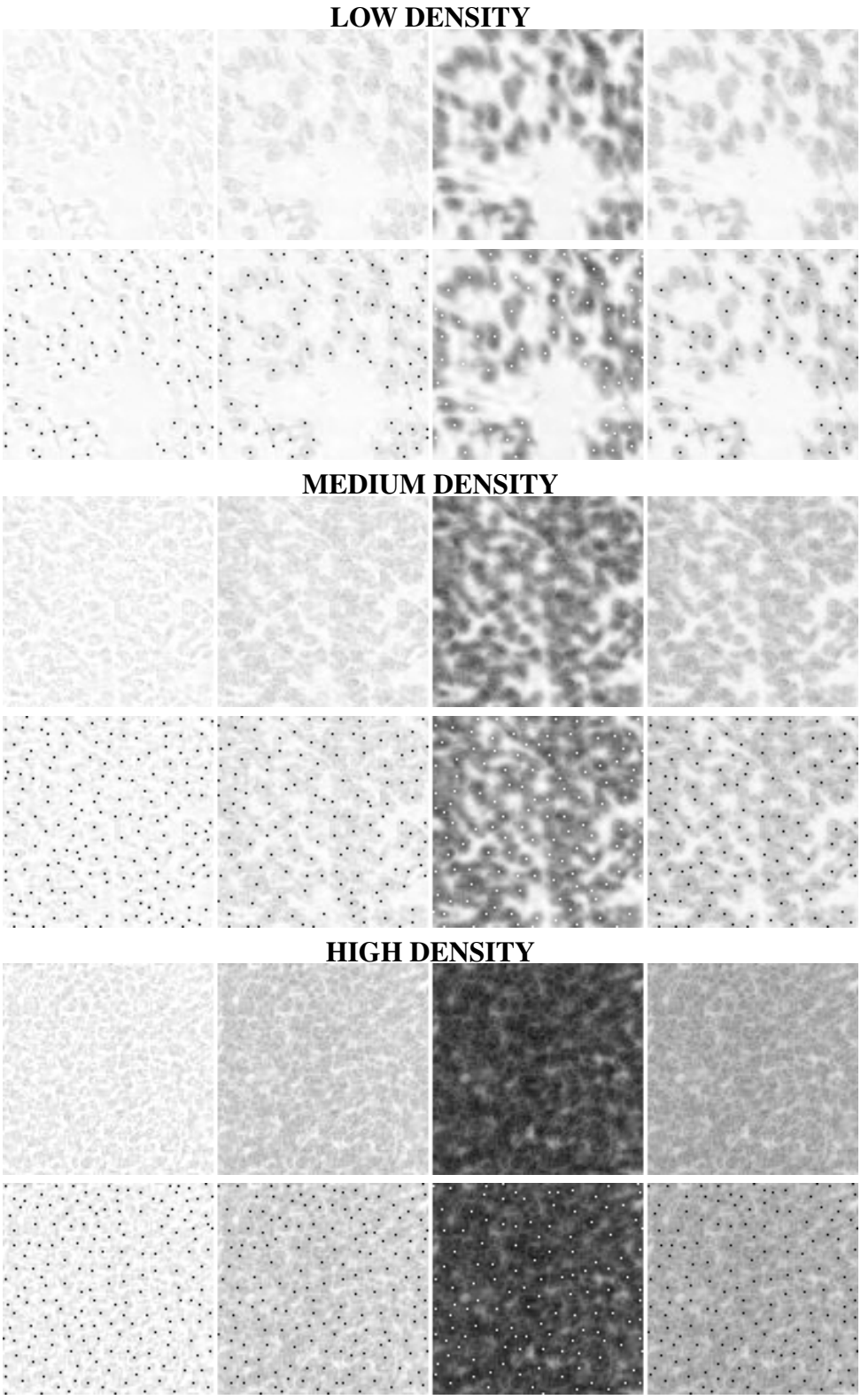


Figure 2.4: Automatic annotation results on data set #1.
First rows: Red, Green and Blue channels and Grayscale. Second rows: combined plot of previous images and the automatically detected filtered points.

A.Density		Low	Medium	High
Avg dist	$m \Rightarrow a$ case	1.82	2.18	2.64
	$a \Rightarrow m$ case	2.77	2.29	2.34
Avg hit %	$m \Rightarrow a$ case	95%	91%	86%
	$a \Rightarrow m$ case	82%	90%	93%

B.Channel	Red	Green	Blue	Gray
Average $d_{m \Rightarrow a}$	2.11	2.07	2.52	2.14
Average std	1.26	1.39	1.70	1.53
Avg Hit %	95%	92%	86%	90%
Wins (#1 or -4 max)	9	5	4	5
Average $d_{a \Rightarrow m}$	2.88	2.37	2.36	2.29
Average std	1.92	1.84	1.83	1.88
Avg Hit %	83%	89%	91%	89%
Wins (#1 or -4 max)	0	9	9	9
Avg dist diff	0.76	0.5	0.67	0.54

Table 2.2: Analysis of results, separated by dataset type (density) and channel.
$d_{X \Rightarrow Y}$ means "distance from X points related to Y". For each point in dataset X the distance to the nearest neighbour in the Y dataset is calculated, or alternatively the point in Y is found that is nearest to X. The values for all datasets are combined to obtain the average and the standard deviation (std) to indicate variation.

2.1.3 Large scale assessment: colour

Studying the method at a larger scale requires using a different approach, since annotating full-sized datasets is impractical and time consuming.

Colour assessment

Typically, density in big images is assessed in a visual manner. Since the counterstain applied paints the cells yellow, the more yellow there is or the darker it looks the denser the area is.

Such an approach is colour-based and, in a sampled version of the image such as the ones used here, would give the value $density = \frac{cellPixels}{totalSpace}$ to each tile, where $totalSpace$ is the size of the tile and $cellPixels$ is the number of pixels that are classified as containing tissue based on their colour. Note that the colour information that will be used is representative of how density is typically examined and gives information about density tendency, but it is far from reliable as will be discussed below.

Automatic *versus* colour

Figures 2.5 and 2.6 link the sampled points obtained via the Hessian-based method and the corresponding colour sampling for the thalamic region in two sections of the reference atlas. Below the original image, the two sampled versions are shown side-by-side.

Overall the reaction to lower density areas such as the corners or ventricle under the midline is similar, with obvious darker areas appearing there. Regarding higher density areas, both images show

very different behaviour. The colour sampling saturates completely around the densest region around the midline, as expected, and any other regions from there contain a decreasing gradient of density. In the point sampling data, on the other hand, the data is much more scattered and does not peak around the midline, resulting in much more detail for the remaining areas.

Further evidence for the saturation of the colour sampling in higher densities is found at the bottom row of the two figures. There, point counts per tile are plotted against their corresponding colour level, first discretely per number with corresponding error bar (left), and then with a cubic regression applied (right) to all the points. Both plots show that point count and colour are correlated quadratically, with a change of curve tendency around the middle density value, 15. It is from there that the colour saturation becomes obvious as the points slowly accumulate at the top end of the plot and reduce in spread.

In the previous section the Hessian-based method was shown to yield best results in cases of medium-high density, so it is likely that the point sampling image is much more representative of the actual underlying tissue than the colour one, even though the latter gives a sense for where the extrema are.

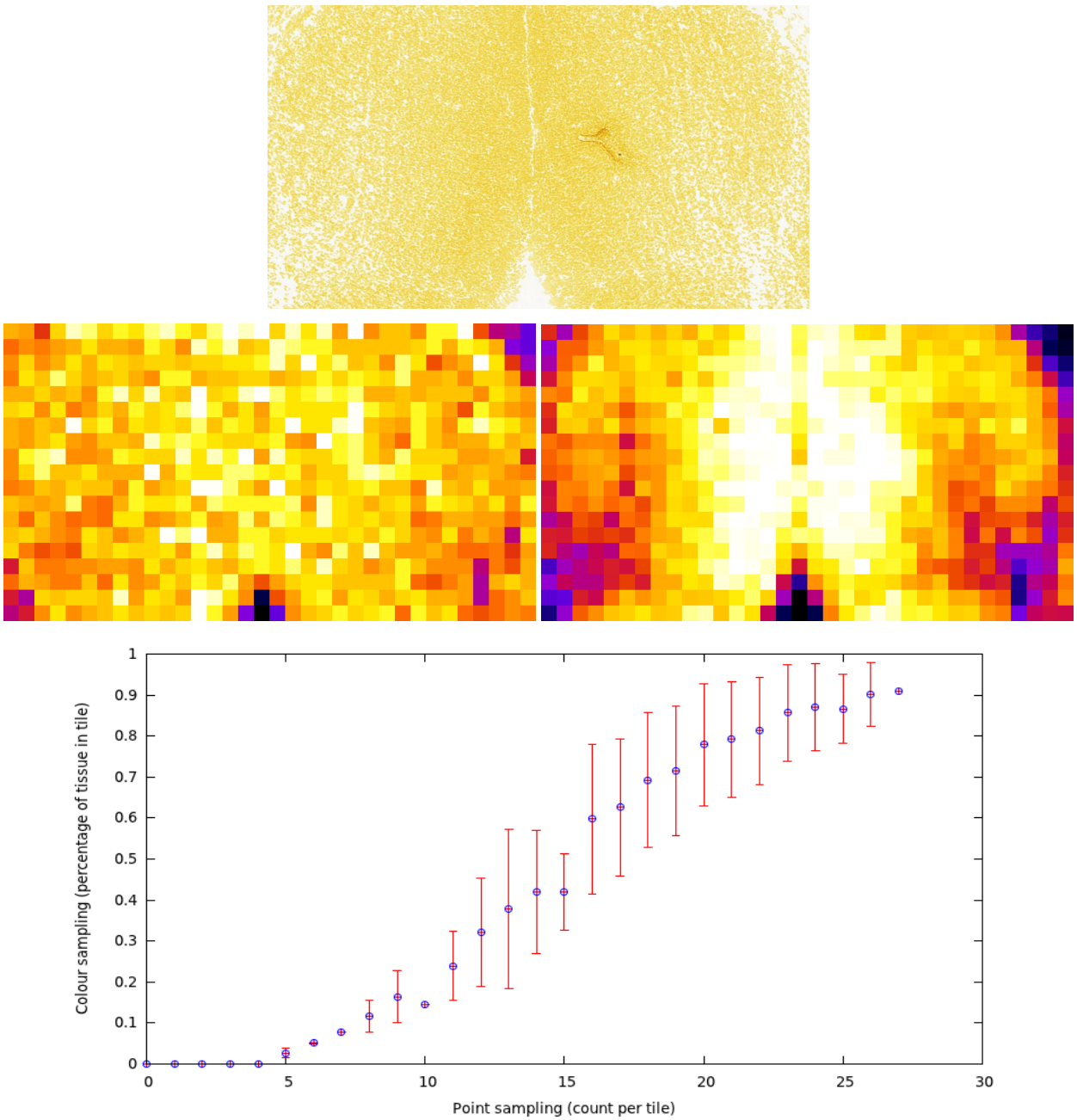


Figure 2.5: Point sampling versus colour sampling for reference atlas image 1. First row: initial counterstained image. Second row: sampled automatically found points and sampled colour. Third row: plot of associated average intensity per point count in tile. Tiles are $40x40\mu m$ squares.

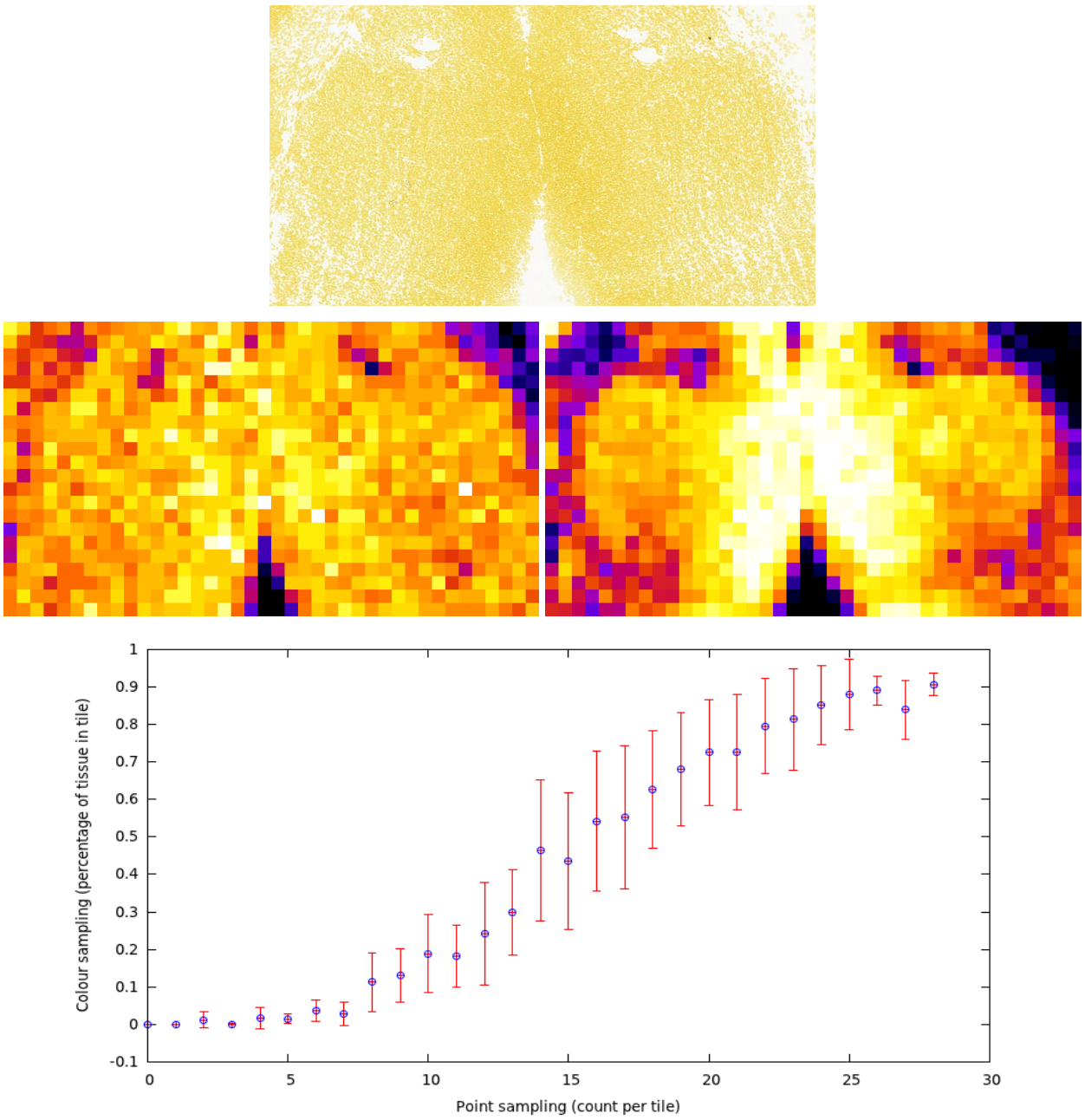


Figure 2.6: Point sampling versus colour sampling for reference atlas image 2. First row: initial counterstained image. Second row: sampled automatically found points and sampled colour. Third row: plot of associated average intensity per point count in tile. Tiles are $40x40\mu m$ squares.

2.1.4 The yellowness case

In the images used to test the method at large scale, the ventricular zone, located next to the midline and the site of cell proliferation, is the region with highest density and colour.

Cells in the ventricular zone are dividing, so some of the nuclei stained will contain twice the DNA. This in turn should mean twice as much counterstain binding, or in other words, twice as *yellow*. Could this situation affect the Hessian-based method and the colour sampling, or would it be possible to reveal whether cells in those areas appear darker?

In figures 2.7 and 2.8 the question is addressed by dividing a reference atlas section into high and low density areas. Cell nuclei are identified with the Hessian-based method, and then the yellowness of each point is obtained from the original image, either as a single value or an average of a circle of radius 3 around the point. This image has been converted to grayscale, where values range from 0 (black) to 255 (white, no tissue).

Next to the original image is a histogram of all the point brightnesses, where the distribution is skewed to the right (white side). To see whether the skewness is due to the two populations of cells, underneath is the split up histogram where high (blues) and lower (red-purple) densities are separated. In the first figure two separated histograms coexist, each one corresponding to radius 0 or 3 being used to obtain brightness associated to a point. In both cases, the resulting distributions are less skewed and it is clear that the high density area contains the darkest cells.

The final plot on the bottom links sampled points and sampled colour similarly to the previous figure 2.5, this time with points coloured according to the density group they belong to. Points in the high density area occupy the upper-right end of the plot, showing the link between automatically measured higher density and darker colour. Figure 2.8 incorporates an additional Medium density category to the previously described analysis, and the three categories do indeed show a distinguishable order from low to high.

If cells in the ventricular zone had twice as much staining and it was apparent by being twice as yellow or ideally twice as dark, the first histograms on the previous two figures would have had two curves, one belonging to the dividing and non-dividing populations. Instead there is a single broader distribution appearing to cover both groups. This could be simply because of the inability of the Hessian-based method to cope with areas with very high density such as proliferative zones. By picking points that are not correct, the presence of this extra staining could be missed.

The staining used could be the problem as well, since it is not clear whether two overlapping dots of yellow will necessarily show a noticeable increase in darkness as a result. Black or gray-coloured staining would be a good approach to test the case and figure out where the problems are.

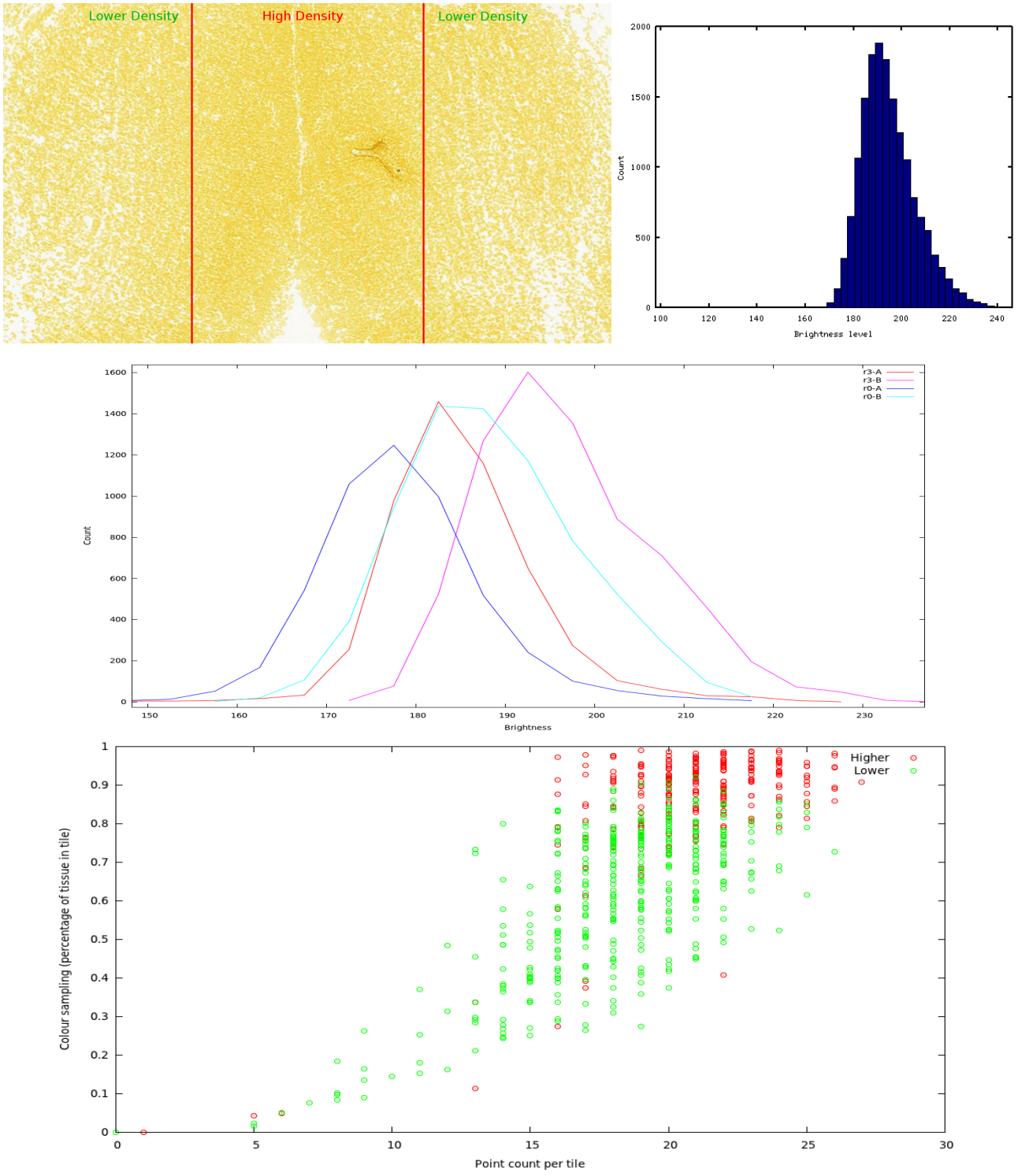


Figure 2.7: Yellowness for the atlas section.
Top row: divisions of the initial image based on visual criteria; Yellowness histogram of the areas surrounding the points found automatically. Bottom row: detailed histogram for the high/low divisions, for radius 0 and 3; Scatter plot of link between sampling of points and of colour, distinguishing between the two density areas.

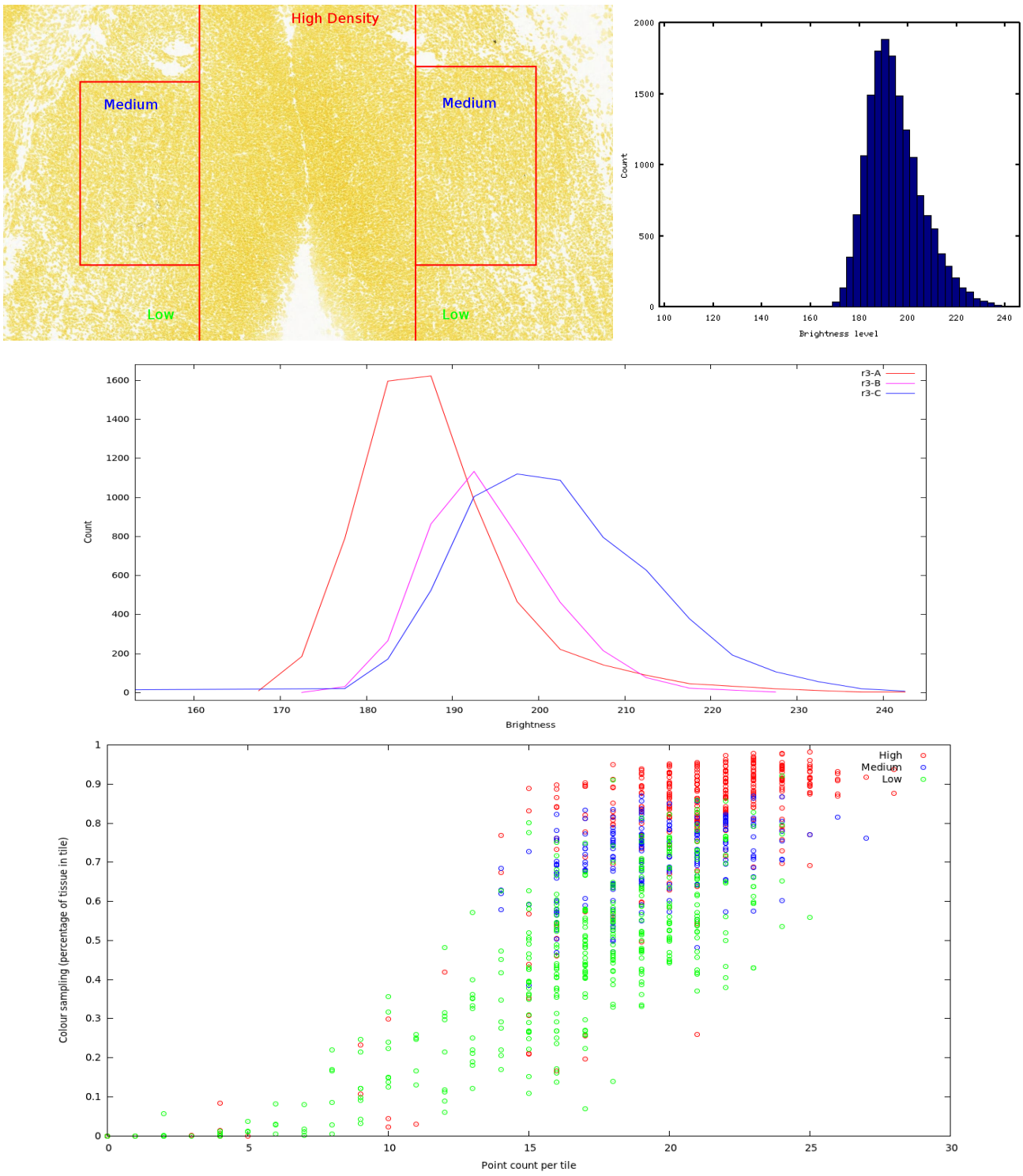


Figure 2.8: Yellowness for the second atlas section.
Top row: divisions of the initial image based on visual criteria; Yellowness histogram of the areas surrounding the points found automatically. Bottom row: detailed histogram for the high/medium/low divisions, for radius 3; Scatter plot of link between sampling of points and of colour, distinguishing between the three density areas.

2.1.5 Conclusions

In the small scale comparison, the Hessian-based procedure yields results that are similar to the manual annotation of the images, as shown by the study of nearest neighbour distance and hit rates between datasets. This similarity in results means that they both agree substantially in their description of the underlying reality, and that they are either representing it reasonably well or being affected by the same shortcomings.

The comparison in the large scale reveals that the proposed procedure does a good job at identifying trends in density. Furthermore, it is less affected by excess accumulation of stained pixels in the densest tissue areas than a basic tile classification based on colour. These results, put together, indicate that the morphological procedure proposed based on the Hessian matrix is a useful tool to extract cell density information from counterstained images in the ADBA.

2.2 Gene expression information

A method has been developed that allows the satisfactory extraction of cell density information from counterstained tissue images. This is the first step towards linking cell density and gene expression. The second is to gather gene expression data from *in situ* experiments.

2.2.1 Utility of Hessian-based method to measure expression

In situ images contain information about gene expression, and because they are counterstained, density. Since the method described previously works with counterstained only images, it would be interesting to see whether this method proves useful to extract density and expression levels from the same image.

The advantage of being able to obtain both values from the same section is that data would be much more accurate. Instead of combining information from two linked *in situ* and reference sections that will be similar but not equal, we would instead have detailed density data from the experimental section on which to build upon to measure gene expression.

The main comparison will be done using two related sections, one from the reference atlas and one from the Neurogenin2 experiment dataset (Fig. 2.9). Both of them are E13.5 coronal sections and have been cropped to focus on thalamus (FR tract on top). Their rotation has been additionally corrected to make the midline central and vertical. The measured area of a cell nucleus is around 60-70μm^2.

In situ and matching atlas section

Visual inspection of the identified points provides a quick way to compare the results of applying the density measuring method to the two matching sections shown in figure 2.9.

Figure 2.10 contains the corresponding results, where a small black disc is drawn in each of the automatically detected points in the images to enable a better inspection of the resulting density. Tissue bordering structures are apparent in both images, specially the axonal tracts that hide the counterstain at the four corners. The medial areas hint at the presence of a midline, but also reveal that the method has some problems when working with *in situ* images: the variation between the two images is readily apparent (the red rectangles in both images show a very affected area). Since this area contains the highest density and the highest level of gene expression either or both of the two factors could be behind this malfunctioning of the procedure.

Problem assessment

To identify the problem a small selection of the midline area was used (figure 2.11). The picture was manually annotated following the same procedure as the previous section (red dots), and the automatic procedure was applied to spot cells (white dots).

Not much overlap between the points can be seen in the overlay picture. That in itself can be explained by the difficulty that a shape detecting procedure has when the borders between the areas to detect are so low and there is so much blurred noise caused by having light (white/gray), nuclear counterstain (yellow) and cell-wide RNA counterstain (black). This is also a highly influential factor for the manual annotation too: it is difficult to be accurate when the two stains are being used.

Some of the automatic points are detected in places that are clearly wrong, such as the midline and other gaps that separate tissue. In the case of counterstained nuclei, these areas would be discarded for being white during the filtering stage. But being a dense area, the presence of many stacked cells -of which some have *in situ* stain everywhere- darkens the whole region and makes the use of thresholds pointless as each area will have differing values.

Conclusion

This inability to account for higher density and higher expression in *in situ* images makes the previous method a poor choice to extract cell density automatically from the same source that provides gene expression information. As a result of this, expression information from *in situ* sections will have to be combined with density information from reference atlas ones. It is not the ideal solution, but having small tissue variation is preferable to rendering density information meaningless.

2.2.2 A colour-based method to measure gene expression

The approach chosen to extract gene expression information uses purely colour information. Based on brightness levels, each pixel in an *in situ* section belongs to one of three categories: no tissue (white, bright), counterstained tissue (yellow, slightly dark) and gene expressing tissue (gray/black, very dark). Because of the automated nature of the data, the values to separate the categories remain constant through different experiments, so they can be used for all the data.

By having counts for these three pixel categories in a specific region, they can be linked to provide meaningful information. In this case, the image is subdivided into tiles to generate data that can be readily matched to the density information obtained previously.

The measure of level of expression for each tile is calculated with

$$expressionlevel = \frac{expression}{tissue} = \frac{expression}{expression + counterstain}$$

where *expression* is the number of pixels that contain gene expression, *counterstain* relates to the pixels with tissue that does not show gene expression, and $tissue$ is the addition of the previous two values. This gives a number for each tile indicating how much of the tissue in it shows *in situ* staining. This method is proportional only to the amount of tissue available and is not affected by density, so it does not include any other factor. Figure 2.12 contains the result of applying this method to the Neurogenin2 section used previously.

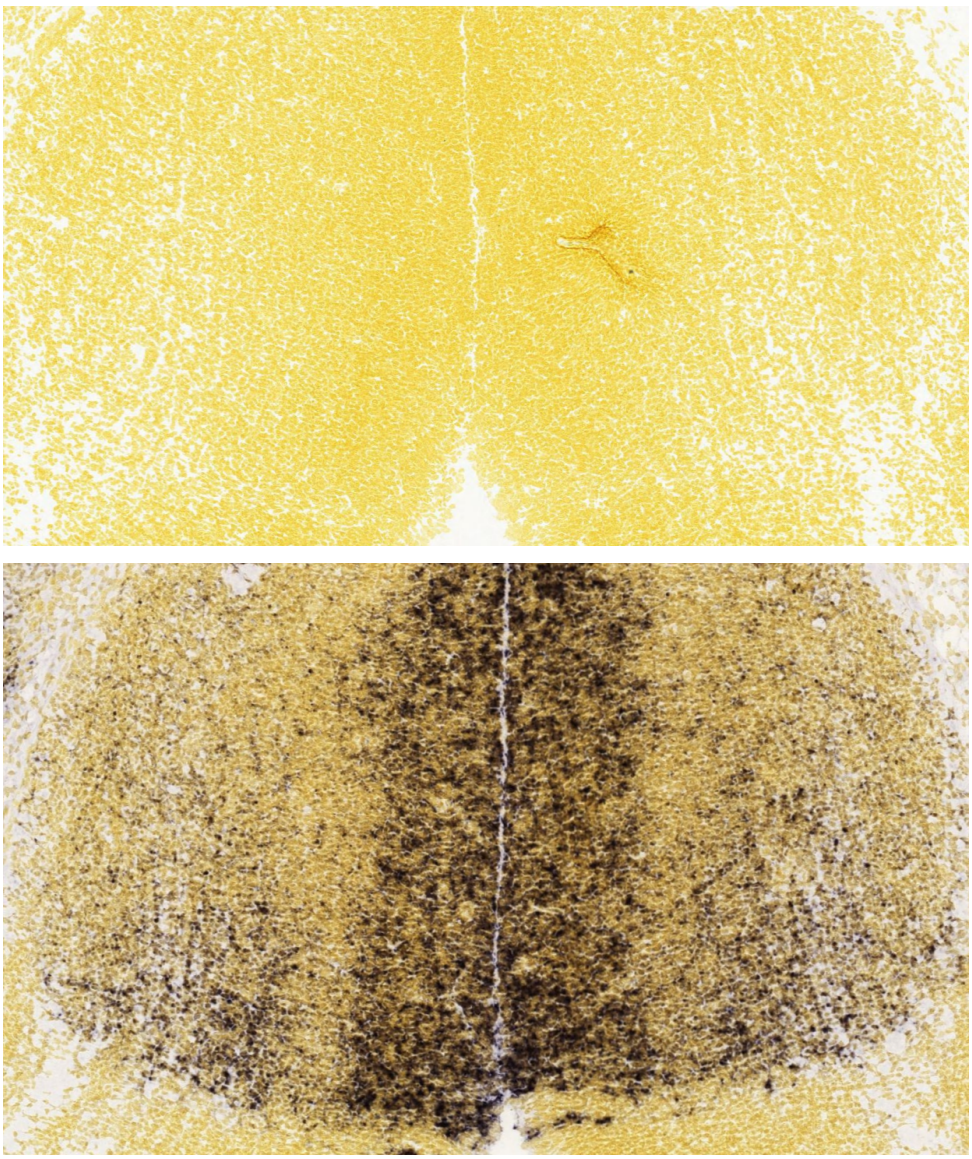


Figure 2.9: E13.5 gene expression sample images.
Original counterstained atlas image and matching similar Ngn2 section to be used to determine whether the cell-detecting method works equally well with gene expression images.

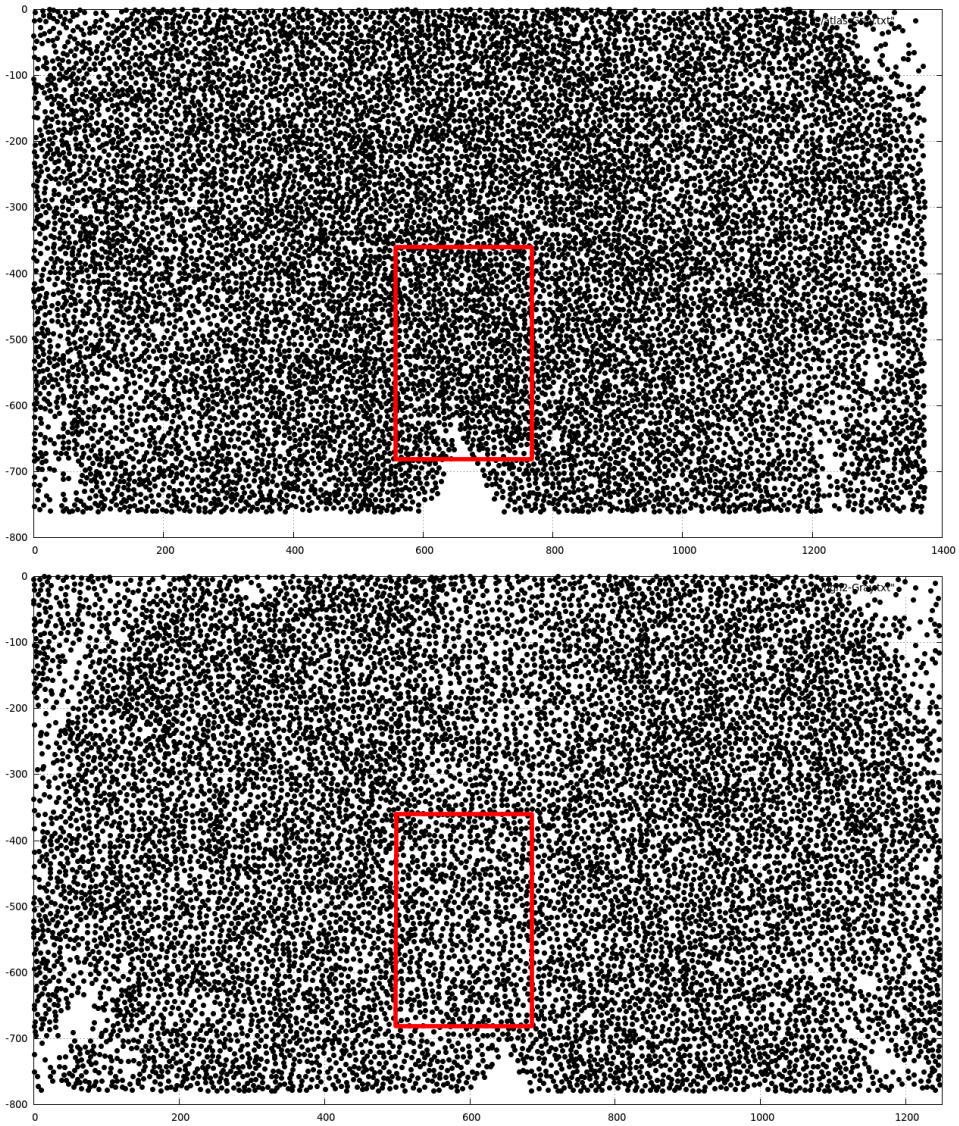


Figure 2.10: Detected points for atlas (top) and Ngn2 (bottom) sections in figure 2.9.

The cell-detecting algorithm has been applied and a simple visual inspection shows that in the case of the Ngn2 section the number of detected points is lower, specially in the midline. The areas in the red rectangle are used to assess what the problems are.

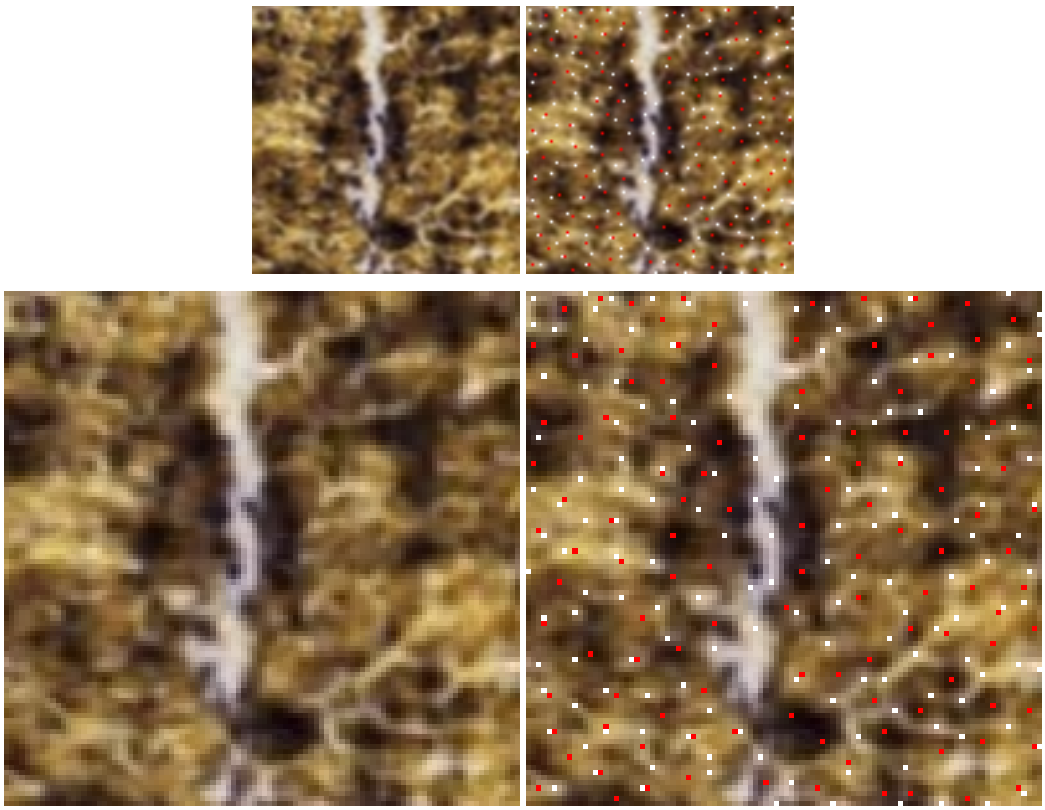


Figure 2.11: Effect of *in situ* staining on the procedure.
Selected area from the midline of the Ngn2 image in figure 2.9 on the left, version with annotated points on the right, and zoomed in (300%) versions below. Red points have been manually annotated while white ones are the result of the automatic procedure.

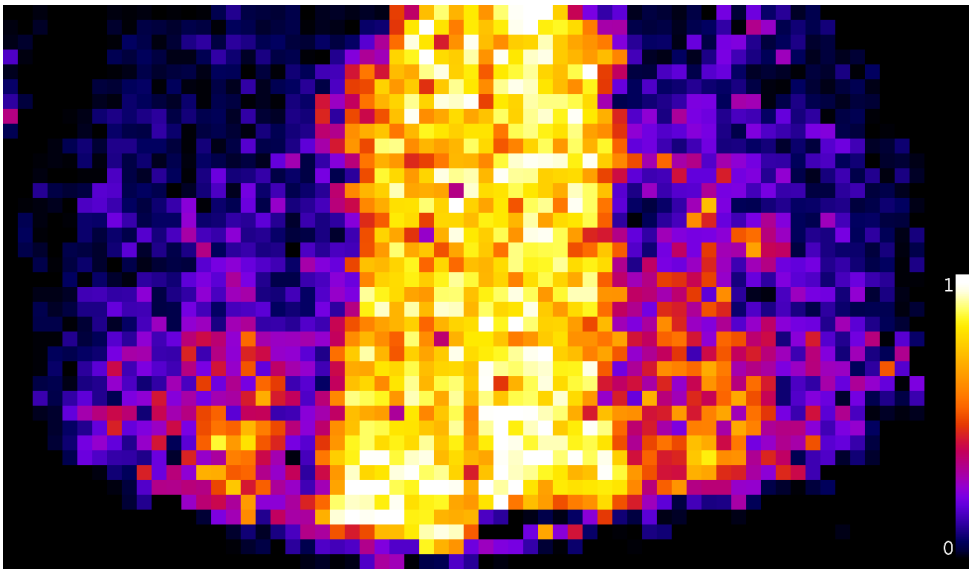


Figure 2.12: Expression levels of the Ngn2 E13.5 section in figure 2.9.
Values range from 0 (none of the tissue expresses the gene) to 1 (all the tissue does).

2.3 Discussion

A method to extract cell location and density of an image based on morphological features has been presented in this chapter.

Its implementation has been discussed, and its results have been tested at a small scale comparing to manually annotated data and at a large scale using colour sampling as a reference. The method has been shown to detect counterstained nuclei reasonably well, with Green and Gray channels being equally useful as source of image information.

The best results are achieved in areas where the density is not extremely high or low, as shown by the cross-examination with the sampled colour. This is not ideal, but it is acceptable for this project, since the final objective is to link this density to gene expression to study double labeling. High and low density areas are readily visible and easy to spot, and are less ambiguous regarding labeling questions. It is in areas with intermediate density where a more accurate measure of it can provide insight on gene expression, and eventually double labeling.

Finally, a method has been proposed to extract gene expression information from *in situ* images. The Hessian-based method could not cope with the kind of data, so a simpler colour-based method has been described.

Chapter 3

Gene co-expression

As a result of the previous chapter, tools are now available that allow the extraction of cell density from counterstained reference data and gene expression from *in situ* experimental sections.

Gene co-expression by groups of cells during development have been shown to be basic for tissue patterning in the first chapter, even up to the point of combinatorially defining the partitioning of space in some cases.

This chapter defines and implements a procedure to measure gene co-expression by combining data generated in the previous chapter. Measurements of two genes are considered at the same time. These can be adapted to any number of genes greater than one, but since the fundamental step is the development of the method it is preferable to test it for a less complex case.

3.1 Data assembly

Data choices and the procedure used for merging them to be used together are now described.

3.1.1 Data used

Reference anatomy and cell density

The source of anatomical and cell density information is the coronal reference dataset for embryonic day 13.5 from the ADBA, described previously.

A total of 55 sections were assembled in an image stack, corresponding to the complete rostro-caudal span of the thalamus. Each of them was first rotated and shifted so that the midline was perfectly vertical in the center, and then aligned vertically to ensure anatomical coherence when traversing the stack. The stack was then cropped to contain only the thalamic area.

Gene expression

Data containing *in situ* hybridization experiments is also available in the ADBA. Following the focus on the thalamus, only genes with detectable expression in the E13.5 thalamus were considered.

The genes chosen are representative of categories that are heavily involved in thalamic development:

- Homeobox transcription factors. As discussed in the first chapter, these are involved in tissue patterning.
  - Gbx2 (Gastrulation Brain Homeobox 2). It is expressed in a subset of thalamic nuclei that require it to differentiate and is a necessary gene for the correct development of the thalamocortical tract [6, 20, 35].
  - Otx2 (Orthodenticle homeobox 2). It is a necessary gene for the correct regionalisation of the diencephalon and the appearance of its structures [32, 26].
- Other transcription factors, which regulate cell identity and development of specific populations among others.
  - Ngn2 (Neurogenin 2). It is expressed in a subset of progenitor cells in the thalamus and is required for the neuron differentiation and patterning of the telencephalon [31, 13], as well as the development of the thalamocortical tract [28].
  - Olig2 (Oligodendrocyte transcription factor 2). It is locally expressed in the developing diencephalon and is required for neuronal differentiation [36, 4].
- Cell adhesion. These are involved in the clustering of cells, basic function during nucleation.
  - Cdh8 (Cadherin 8). It is expressed in a localised manner in the diencephalon and telencephalon and it is related to nucleation and tract development [33].
- Cell signaling. EphA4 is a receptor for signaling molecules, involved in axon navigation and the establishment of ordered connections
  - EphA4 (Ephrin receptor A4). With its ligand ephrin-A5, they are expressed locally in thalamic nuclei and target structures as gradients and are involved in tract formation and the forming of orderly connections as maps [11, 9, 37].

	Gbx2	Ngn2	Olig2	Otx2	Cdh8
Ngn2	3				
Olig2	2	0			
Otx2	3	1	2		
Cdh8	4	2	2	5	
EphA4	3	6	2	4	4

Table 3.1: Number of comparable sections for each pair of genes.

In the 55 sections of the reference atlas that contain the thalamus, the sections of every pair of genes are comparable at least once except for Ngn2-Olig2 (table 3.1 contains the full details). Genes that will be compared should ideally have several comparable sections and be distributed across the thalamus: 14 pairs of gene expression are therefore usable for further comparison with the procedures discussed next.

3.1.2 Gene expression integration

Data corresponding to *in situ* experiments undergo the same pre-processing as for the reference set. The sections containing the thalamus are rotated and shifted so that the midline is vertical to allow better comparison on the next step.

Each of the sections containing gene expression is then matched to one in the reference dataset that looks most similar. The choice is made using visual observation of shape, size and tracts (fasciculus retroflexus, thalamocortical, and mammillothalamic rostrally), and always taking into account the variations in slicing angle and inter-individual anatomy. While the procedure is not perfect, it yields results that reasonably transfer features from one image to the other.

To assemble the combination of each density-expression pair together, the areas showing gene expression from one image are selected, copied and then pasted into the other (see figure 3.1). This *virtual grafting* procedure has as a result the incorporation of gene expression into a system where cell density and anatomy are known. As a final step, these combined images have their expression sampled via the approach described in the previous chapter, yielding images ready to be used.

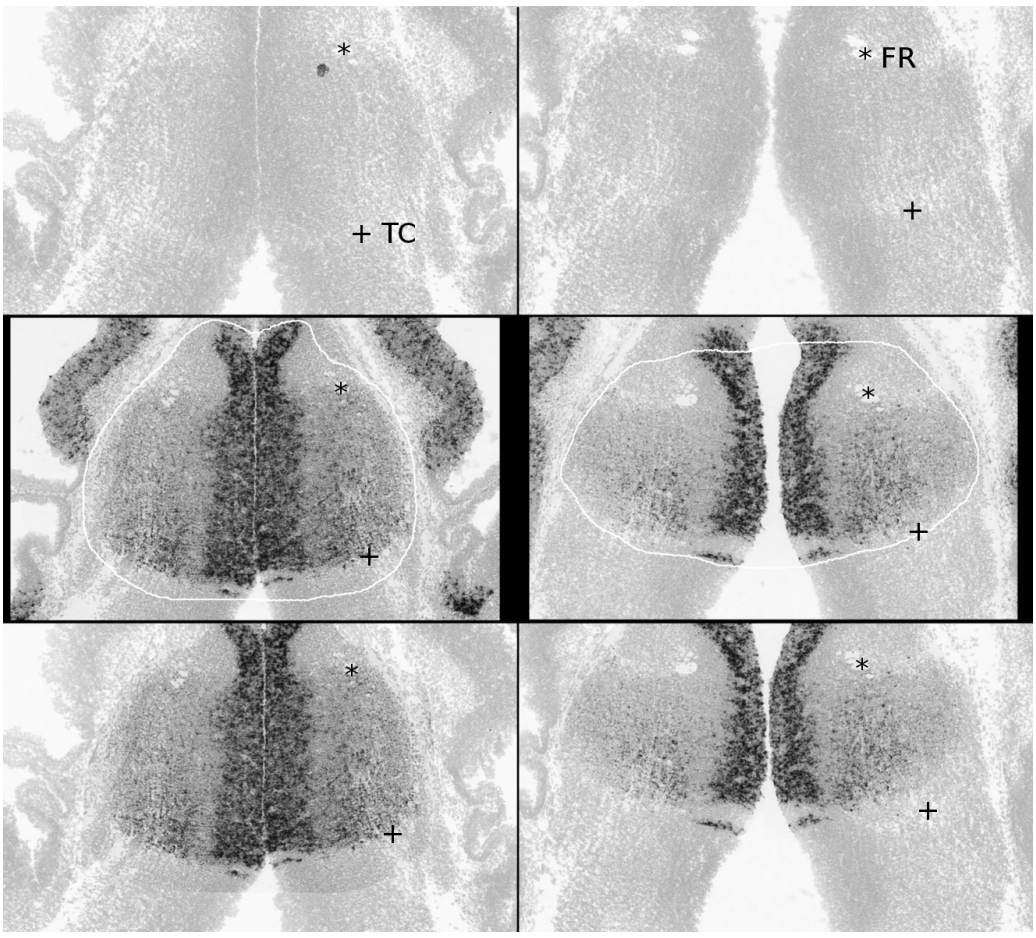


Figure 3.1: Virtual grafting procedure.
A reference atlas section (top row) is matched to an experimental *in situ* section (middle row, Ngn2 in the two examples). The area of gene expression, delimited by white, is copied and pasted to the reference image to obtain an image that combines both (bottom row). *Asterisk* (*): Fasciculus Retroflexus, perpendicular axon tract marking the upper thalamic limit. *Plus sign* (+): Thalamocortical Tract.

3.2 Double labeling calculation

Information about cell density and gene expression has been gathered and a method has been described to link them inside the anatomy of the thalamus. The next stage is to extract useful data about gene co-expression and to interpret it.

3.2.1 Gene frequency maps

The integration procedure described in the previous section generated, for each experimental section revealing gene expression, a section of the reference atlas where that expression had been *grafted* and then sampled. This means that several of the reference atlas sections have been divided in tiles and for each tile we know how much of the tissue expresses a specific gene.

Given two genes that have sections in comparable locations, it is possible to link the two levels of expression to gather knowledge about their relationship. In the specific case of this project, and as shown in table 3.1, all but one pair of genes can be compared in at least one section, so 14 interactions can be analysed.

Two measurements describing levels of gene co-expression can be readily obtained by combining the two separate expression samplings and are described next.

Minimum double labeling

The amount of minimum double labeling tells us whether two genes do *necessarily* overlap in a given tile, and how big that overlap is.

For the distributions from the two genes to guaranteed to have some overlap, the percentages of tissue they cover per tile must add up to more than 100%. For example, if the two genes cover 10% and 20% of the tissue respectively, they may have no overlap; if they have 60% and 70%, they must have some overlap. If expr(X) denotes the fraction of the tile covered by the expression of gene X, the minimum amount of double labeling is

$$min2x(A,B) = max(\,0\,, expr(A) + expr(B) - 1\,)$$

Maximum double labeling

The minimum double labeling is a conservative measure that reveals necessary overlaps. In contrast, the maximum double labeling gives a measure of how much of the tissue *could* be expressing both genes at the same time.

The calculation yields as a result the minimum amount of tissue expressing either of the genes, since this could be the maximum amount happening to co-express the genes. The formula used to calculate the measure is

$$max2x(A,B) = min(\,expr(A), expr(B)\,)$$

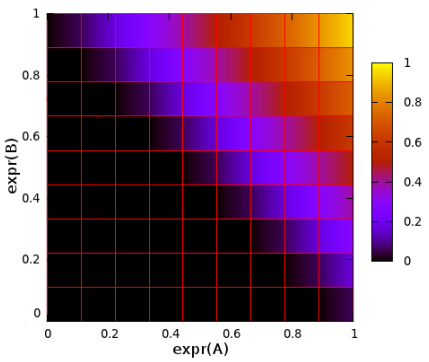


Figure 3.2: Minimum double labeling depending on the expression of genes A and B.

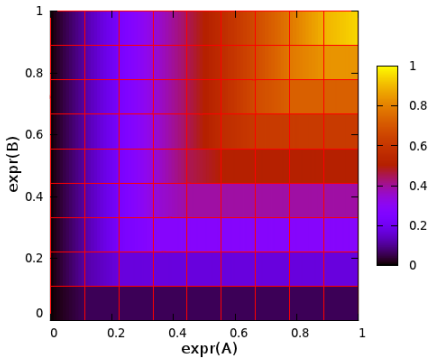


Figure 3.3: Maximum double labeling depending on the expression of genes A and B.

Application

The combination of both measurements offers realistic information about the range of double labeling expectable between two genes that are expressed in a comparable tissue section. It is only when the theoretical double labeling frequencies are multiplied with the measured number of cells for each tile that the measured number of cells that do co-express is obtained. This cell count, in turn, is meaningful and has practical uses as opposed to labeling frequencies.

Figure 3.4 shows the results of applying the two measurements to the sample sections of two genes. In some tiles the range effectively moves from 0 to the described maximum double labeling, while in others the range is much smaller and it indicates less variability of the calculations.

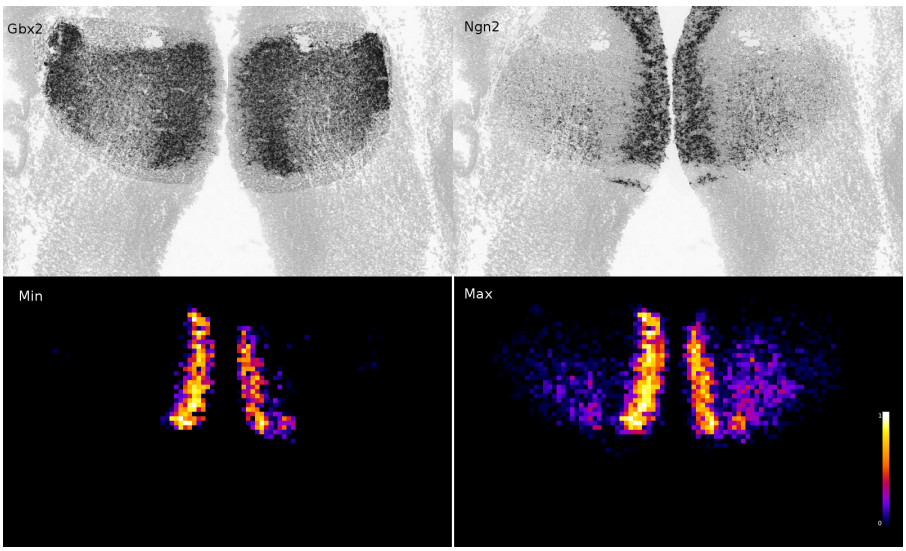


Figure 3.4: Amount of double labeling.
Calculation of minimum and maximum double labeling (bottom left and right respectively) for a specific section of Gbx2 and Ngn2. The values indicate fractions and the range is therefore [0,1].

3.2.2 Calculation of cell numbers by incorporating density

The calculation of maximum and minimum double labeling has been based on the sampled gene expression which in turn relies on colour information about the *in situ* stain.

Since we have calculated the density of every single section in the reference atlas, and because the gene expression has been integrated in its anatomical space, the two can be combined. As a result we no longer need to consider theoretical percentages of tissue, but can actually talk about *specific* numbers of cells. Additionally, as the tile size is $20x20\mu m^2$ and one tile contains a maximum of 11 cells, it is also interesting to see how many cells the calculated percentage corresponds to.

To calculate the density-normalised labeling figure it is only necessary to multiply it with the density from the corresponding reference atlas section containing the sampled density. Figure 3.5 shows examples of this procedure and the effects it has. Note that the background and colour range values are opposite to those used in 3.4 to make it easy to identify minimum/maximum amounts of double labeling or their application to obtain density numbers.

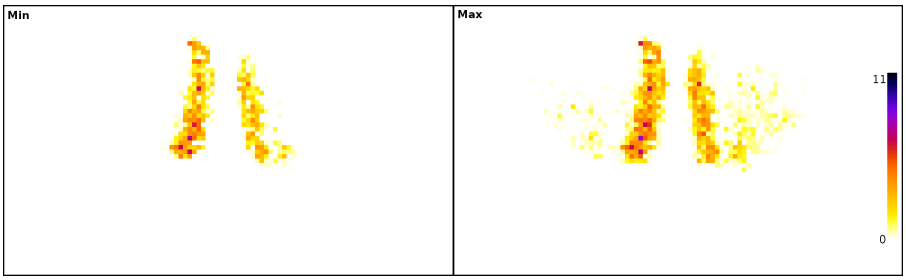


Figure 3.5: Density application.
Effect of applying density to sample results in figure 3.4.

3.2.3 Data interpretation

The combination of minimum and maximum double labeling values for each pair of genes that have been corrected for density provides information of the number of cells that *must* be co-expressing both genes and the potential number of cells that *can*. Additionally, the information is available per tile, so it is possible to analyse variation over anatomy or quantify areas of interest.

Dividing the values of minimum and maximum double labeling according to whether they are zero or higher, four combinations appear that provide a way to classify the level of co-expression. Since the minimum double labeling can never be higher than the maximum, there are three realistic co-expression possibilities, defining three classes as shown in table 3.2:

		Min. DL	
		0	>0
Max. DL	0	No co-expression (C)	-
	>0	Possible co-expression (B)	Co-expression (A)

Table 3.2: Gene co-expression cases.
Possible combinations of minimum and maximum double labeled cells and their implications on gene co-expression. These can be studied at the tile, section or dataset level.

There are different levels at which these combinations can be used to define relationships. For example, it is possible to compare two tiles to obtain very localised information. The full images could be also considered and, building on tile comparison, a general relationship could be defined for the two genes given the section. More generally, it would be possible to define the co-expression relationship between two genes by incorporating the results from all the sections they share. This level describing the dataset is the most appropriate for this project since we are studying pairs of genes.

Three classes of gene co-expression can be therefore defined to categorise each gene pair based on the possible combinations: A, B and C. The class of a given pair of genes will depend on the highest class of any of the sections they have in common: even if two genes only show necessary co-expression in one of the five sections they share, the class of the pair will still be A (co-expression).

Class C indicates there is no double labeling at all. In the case of this project, since all the genes chosen do overlap even minimally, this class is theoretical and the data does not contain any examples of it.

Class B comprises the cases where the level of co-expression ranges from zero to a value greater than zero (figure 3.6). The possibility of co-expression exists but it is not guaranteed.

Class A indicates where both minimum and maximum values are >0 (figure 3.7). Here co-expression does happen, and the difference between the values describes the range of variation.

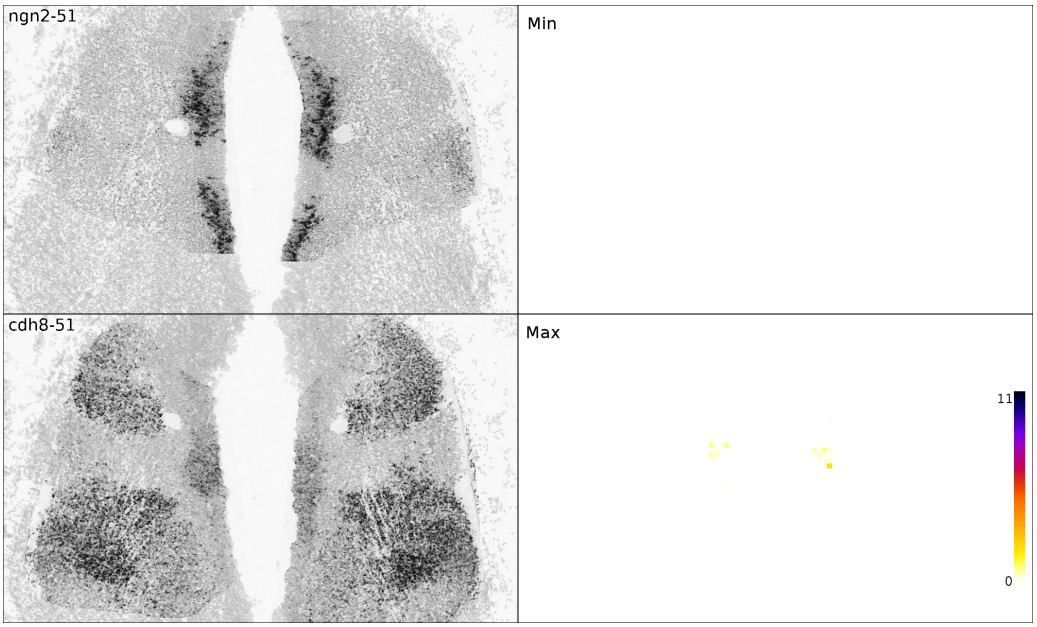


Figure 3.6: Sample of a class B gene co-expression situation, where min=0 and max>0. Left column: expression of both genes. Right column: minimum (top) and maximum double labeling (bottom).

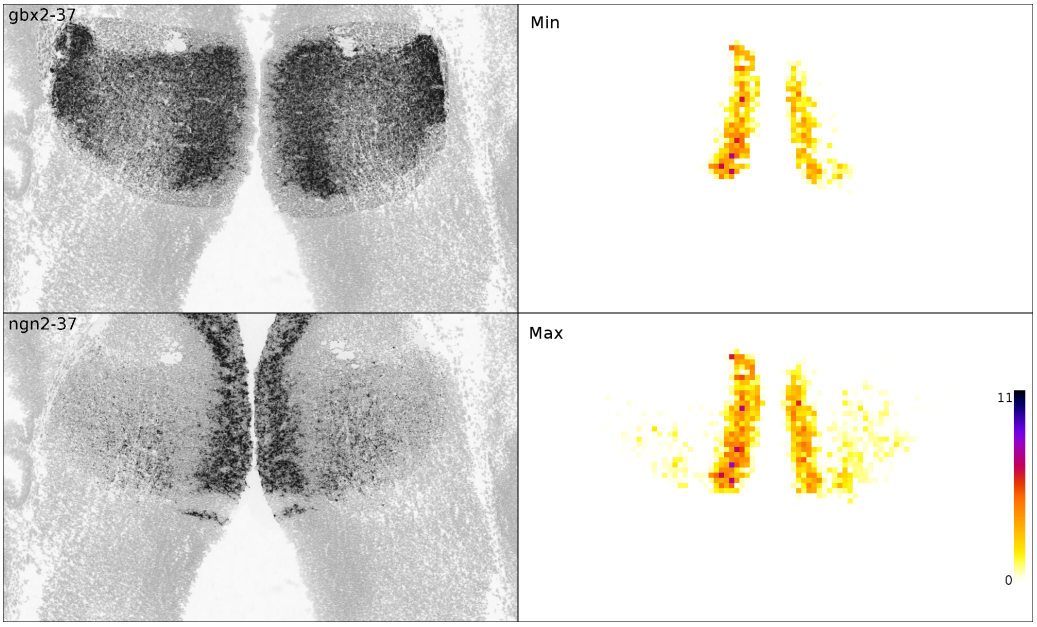


Figure 3.7: Sample of a class A gene co-expression situation, where min>0 and max>0. Left column: expression of both genes. Right column: minimum (top) and maximum double labeling (bottom).

3.3 Results

3.3.1 Gene relationship classifications

In the previous section three classes were defined to establish the type of co-expression two genes had in a specific section. The next step is to assign a class to a pair of genes based on all their comparable sections.

To do so, a strategy based on the class of each section of the pair is used. First of all the maximum value in the minimum and maximum number of double-labeled cells are calculated for all the sections (results are in table 3.3). For example, in the case Ngn2-Cdh8, the first section has minimum and maximum values of 5 and 5 cells, while for the second one the numbers are 0 and 2 cells.

Using the values for the table above, the class of each section is then calculated. All the sections from a gene pair are combined to obtain the class of their co-expression relationship. Note that the highest section class determines the pair class. An additional plus sign (+) is used to mark those genes that not only are minimally co-expressed somewhere in the thalamus, but actually do so in all the sections. Table 3.4 contains the resulting classification.

Gbx2	-	770/781	00/12	347/467	5321/6433	040/151
Ngn2	770/781	-	N/A	5/5	50/52	003100/123212
Olig2	00/12	N/A	-	25/26	02/01	00/02
Otx2	347/467	5/5	25/26	-	24535/45646	0000/1111
Cdh8	5321/6433	50/52	02/01	24535/45646	-	0000/1121
EphA4	040/151	003100/123212	00/02	0000/1111	0000/1121	-
	Gbx2	Ngn2	Olig2	Otx2	Cdh8	EphA4

Table 3.3: Maximum minimum and maximum double labeling values.

Gbx2	-	A aab	B bb	A+ aaa	A+ aaaa	A bab
Ngn2	A aab	-	N/A	A+ a	A ab	A bbaabb
Olig2	B bb	N/A	-	A+ aa	B cb	B cb
Otx2	A+ aaa	A+ a	A+ aa	-	A+ aaaaa	B bbbb
Cdh8	A+ aaaa	A ab	B cb	A+ aaaaa	-	B bbbb
EphA4	A bab	A bbaabb	B cb	B bbbb	B bbbb	-
	Gbx2	Ngn2	Olig2	Otx2	Cdh8	EphA4

Table 3.4: Classification of the level of co-expression between gene pairs.
For each gene pair, the class of each of their common sections is shown in lowercase on the lower part of the cell. On top of them, the resulting gene co-expression relationship class is shown in capital letters.

3.3.2 Result grouping

The 55 section span of the thalamus has been divided in four regions: rostral/caudal ends, corresponding to the 10 sections at each end, and rostral/caudal, dividing the middle region evenly (reference images are shown in figure 3.8). The reason for this subdivision is to allow the comparison between different gene pairs given that the number of sections they can be compared on are sometimes too few and not too scattered.

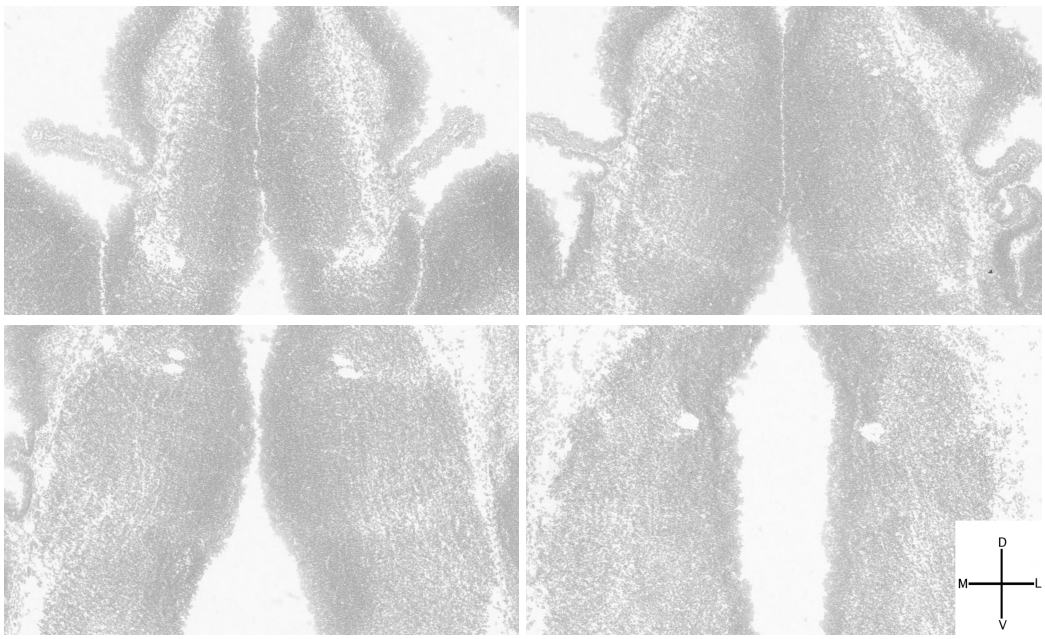


Figure 3.8: Source images for the four subdivisions.
Top row: rostral end, rostral. *Bottom row*: caudal, caudal end. Note that all the subsequent figures use the same axes, where the midline is exactly in the center, the lateral ends are on each side, and the dorso-ventral axis runs vertically.

3.3.3 Figures

The corresponding figures for all the 14 gene pairs studied in this project are now ordered based on their overall class. Within each class the order relates to the number of sections providing evidence of belonging in the class they contain: those pairs containing lowest number of sections are shown first.

Each section is a column made of four pictures. The top two correspond to the expression of the two genes, and are labeled with the gene name and number of section in the thalamic subset. The bottom two images contain the minimum and maximum amounts of double labeling. Columns are sorted rostrally (left) to caudally (right).

Because all the chosen genes are expressed in the thalamus at E13.5 and they all have minimal overlap, no pair of genes belongs to Class C. Nevertheless, as shown in table 3.4, two pairs of sections do belong to this class. There are many gene pairs that also have a class C relationship in big areas in the thalamus, as a consequence of the gene expression overlaps being minimal.

Note that in some cases the measurements of minimum and maximum co-expression are very faint, and as a result it can be hard to examine the figures visually. This effect becomes more apparent the lower the class of a gene pair is.

Table 3.5 gives details of which gene pair may be expressed (class B) or is co-expressed (class A) and a description of the co-expression.

	Location of co-expression in the thalamus
Class B	
Olig2 EphA4	*Ventro-medial area of the caudal end*
Gbx2 Olig2	*Ventro-medial area*
Olig2 Cdh8	*Ventro-medial area*
Cdh8 EphA4	*Lateral areas*
Otx2 EphA4	*Medial and ventral rostrally, dorso-medial caudally*
Class A	
Ngn2 Cdh8	Small ventro-medial cluster at the caudal end
Gbx2 EphA4	Medial and lateral sides *Most of the thalamus, dorso-lateral cluster*
Gbx2 Ngn2	Big central vertical band rostrally, then concentrated on the medial area *Scatter over the thalamus caudally to dorso-lateral clustering*
Ngn2 EphA4	Small dorso-medial area *Wide scatter throughout that concentrates on medial and lateral ends at the caudal end*
Class A+	
Ngn2 Otx2	Medial band that thickens ventrally
Olig2 Otx2	Ventro-medial cluster rostrally, dorso-medial cluster at the caudal end
Gbx2 Otx2	Central vertical band at the rostral end that becomes a thicker medial band caudally
Gbx2 Cdh8	Thick central vertical band that becomes medial *Wide dorso-lateral scatter that concentrates on its corner at the caudal end*
Otx2 Cdh8	Thick medial band with a strong cluster at the ventro-medial end *Ventral band that becomes more intense caudally*

Table 3.5: Co-expression description for the gene pairs.
The type of co-expression is indicated by the emphasis used: regular text indicates definite co-expression, while *italics* indicate it is possible.

Class A+

Ngn2-Otx2

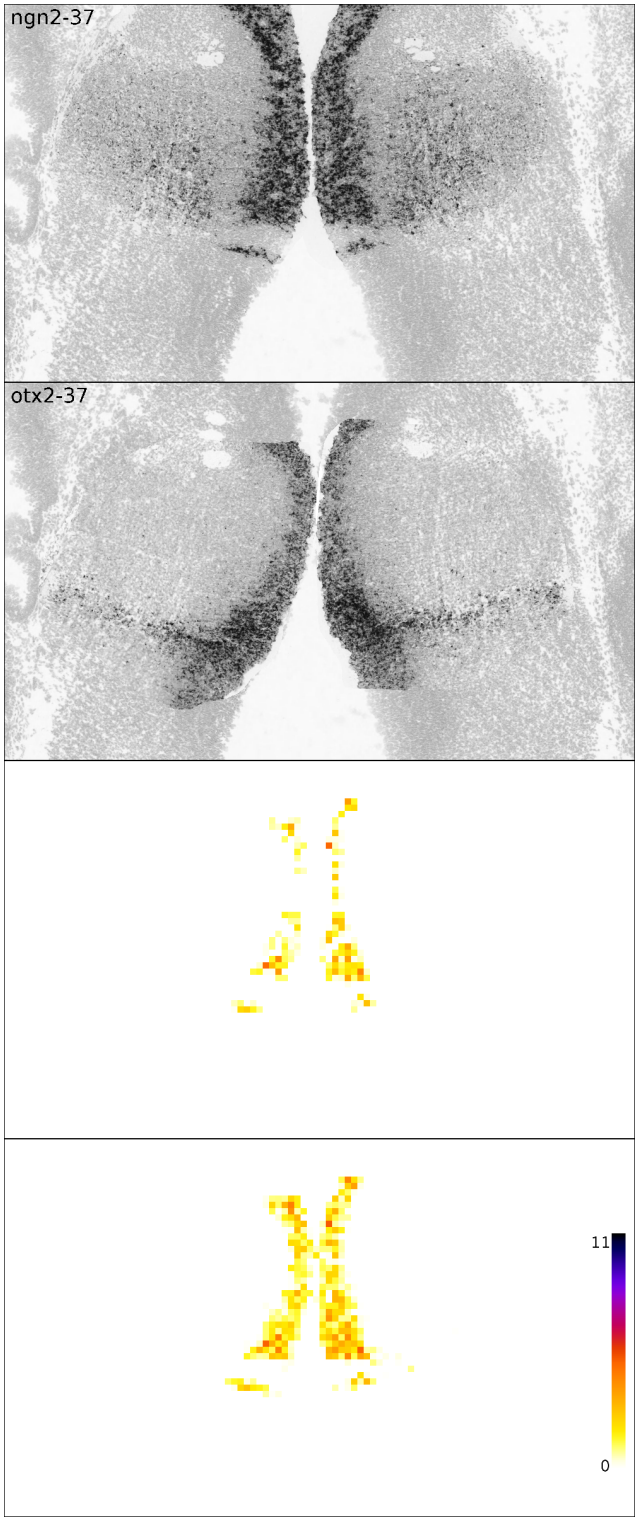


Figure 3.9: Ngn2-Otx2. Caudal data. Single overlapping section.

Co-expression, both potential and certain, is located at the medial end of the thalamus. The clusters at the ventromedial corner appear in both the minimum and maximum double labeling figures so that indicates that co-expression does happen. In contrast the dorsomedial corner contains differences between the figures, suggesting a only potential co-expression. This data interpretation is used throughout the series of figures and will be aggregated in the following section.

Olig2-Otx2

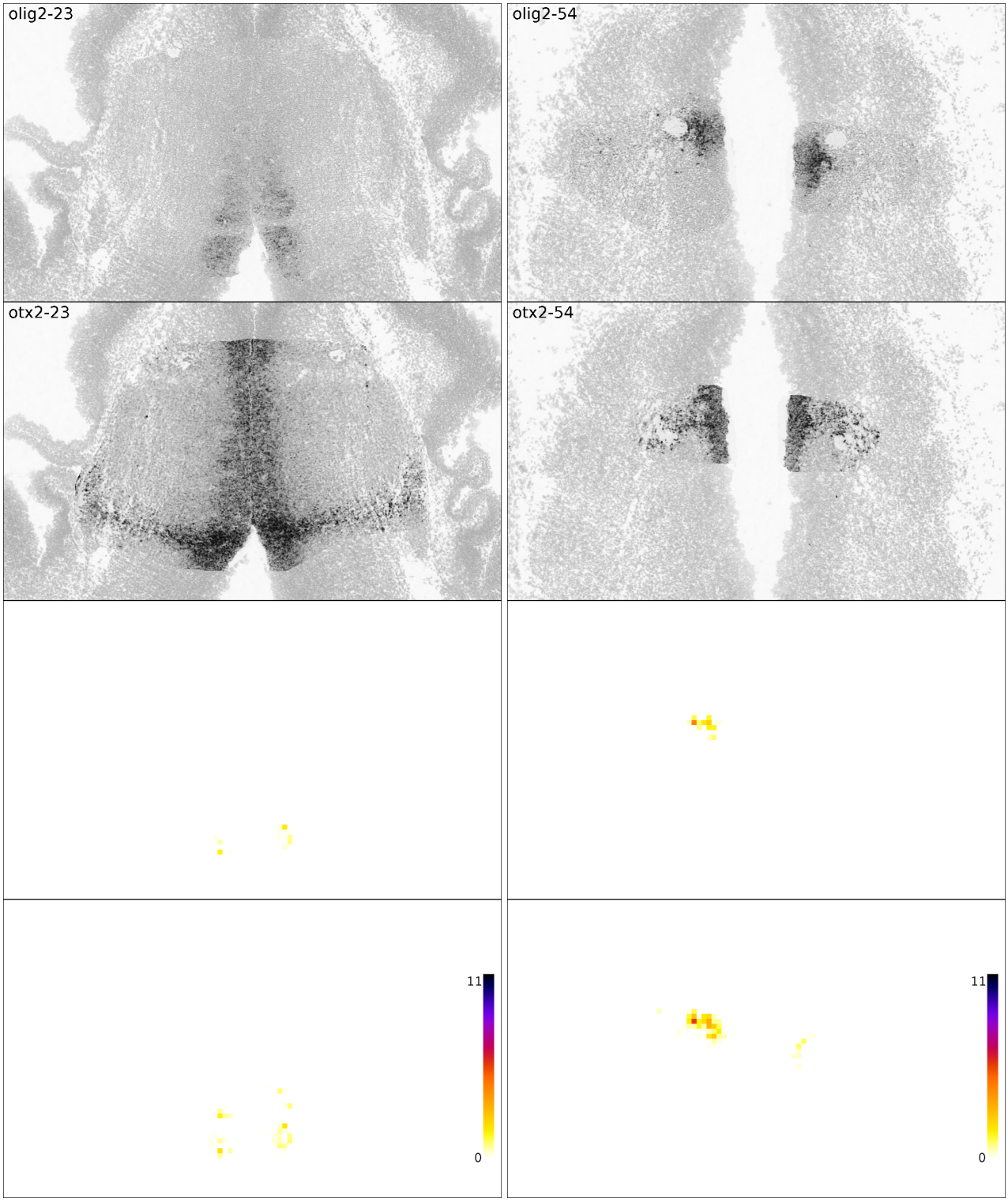


Figure 3.10: Olig2-Otx2. Rostral (left) and caudal end (right) data.

Gbx2-Otx2

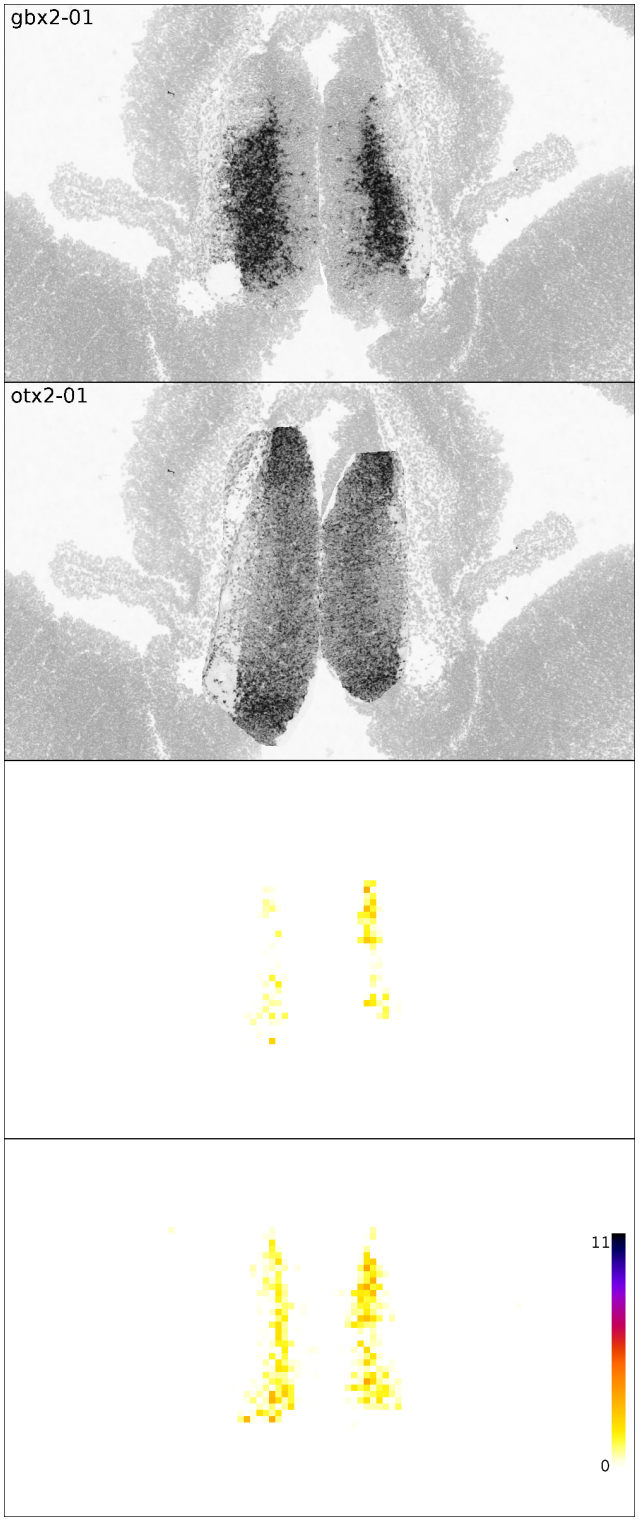


Figure 3.11: Gbx2-Otx2. Rostral data.

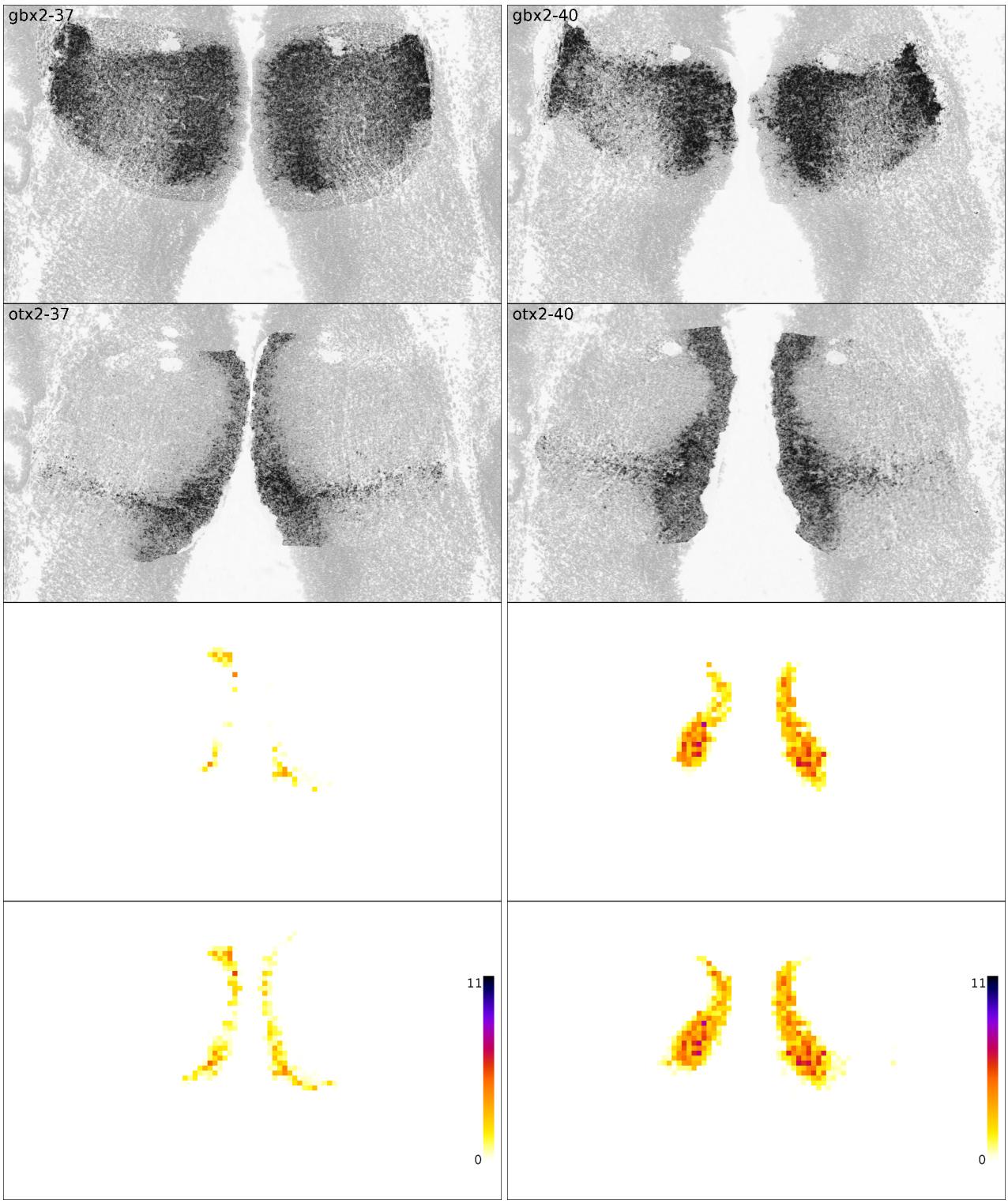


Figure 3.12: Gbx2-Otx2. Caudal data.

Gbx2-Cdh8

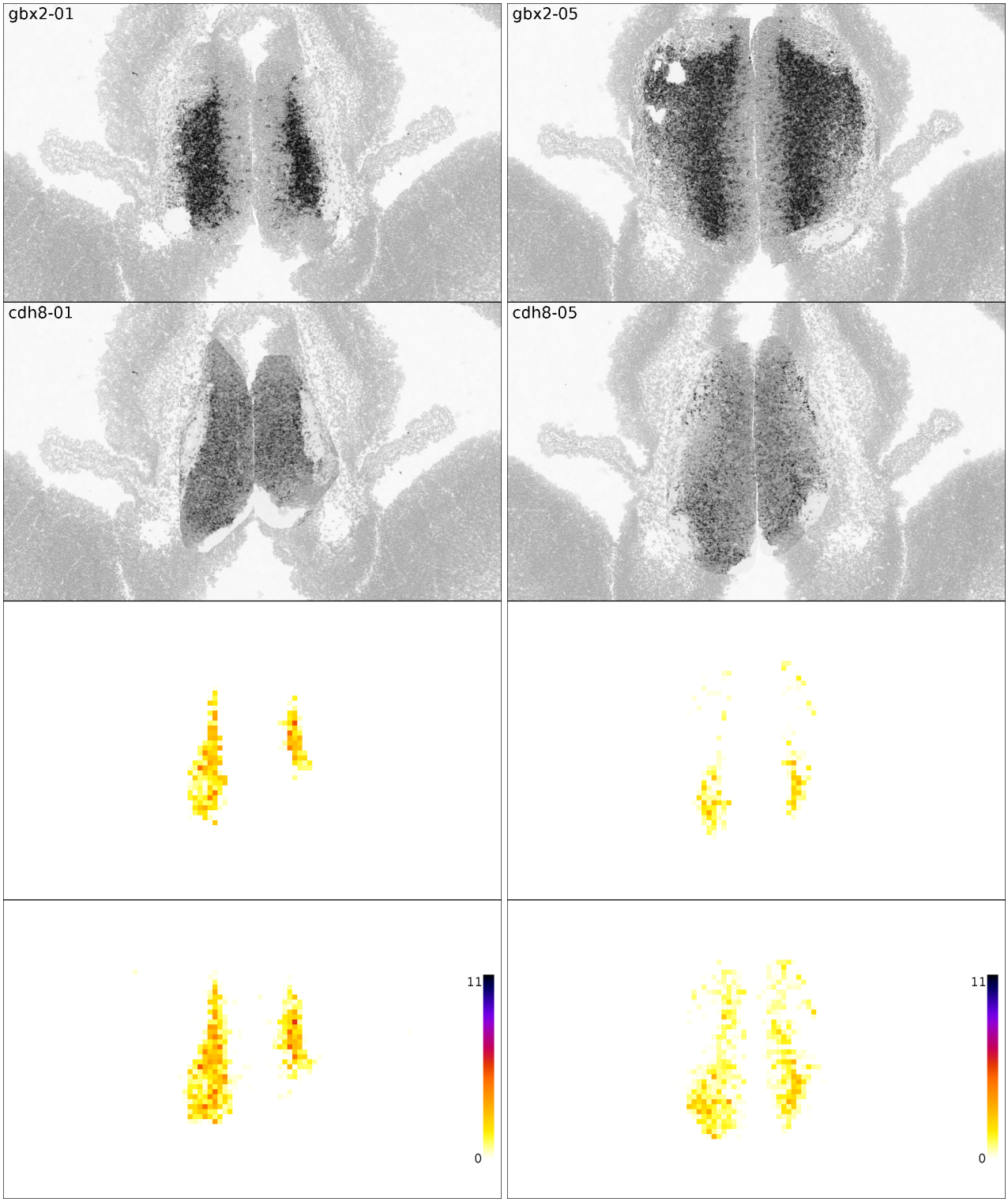


Figure 3.13: Gbx2-Cdh8. Rostral end data.

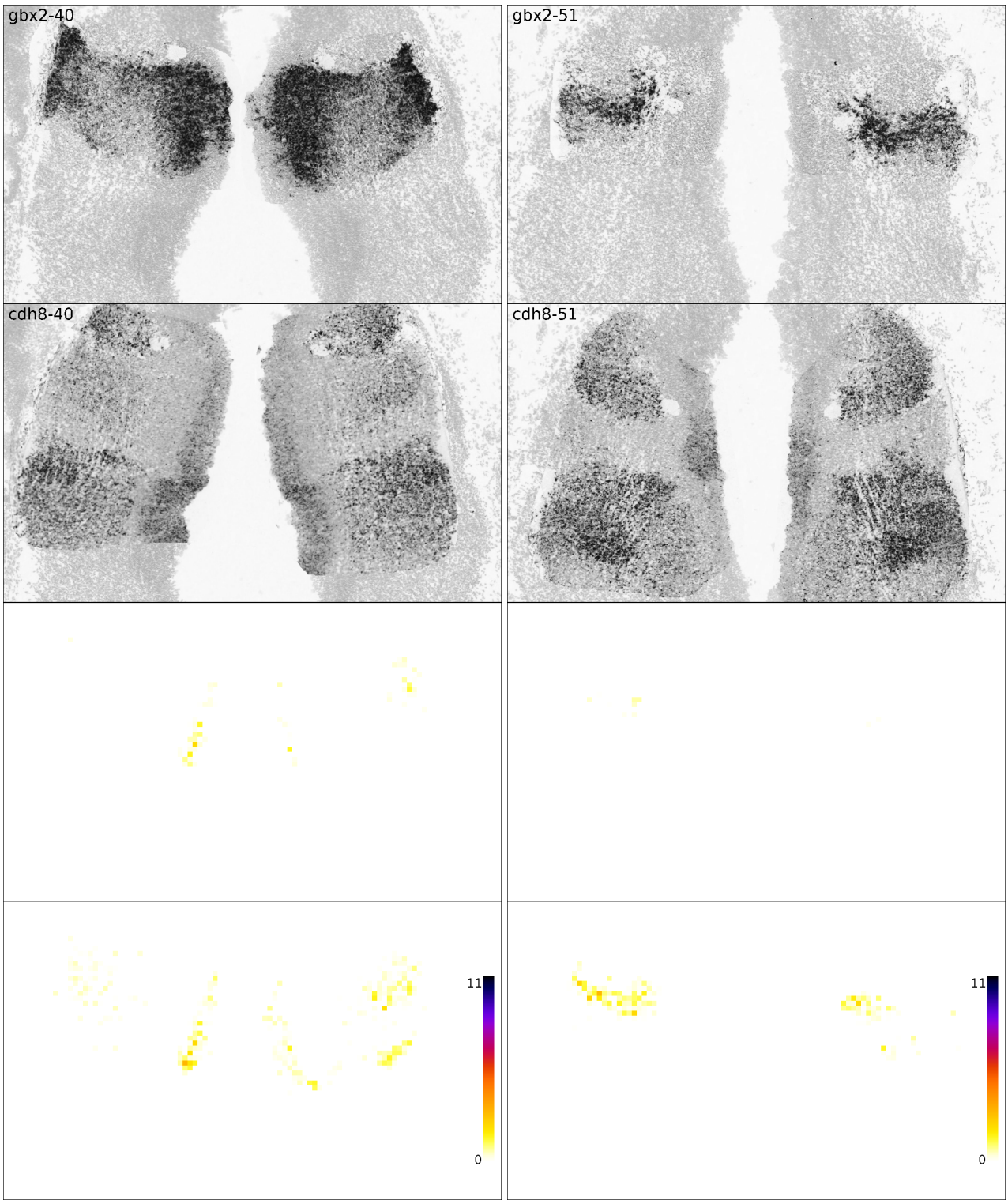


Figure 3.14: Gbx2-Cdh8. Caudal (left) and caudal end (right) data.

Otx2-Cdh8

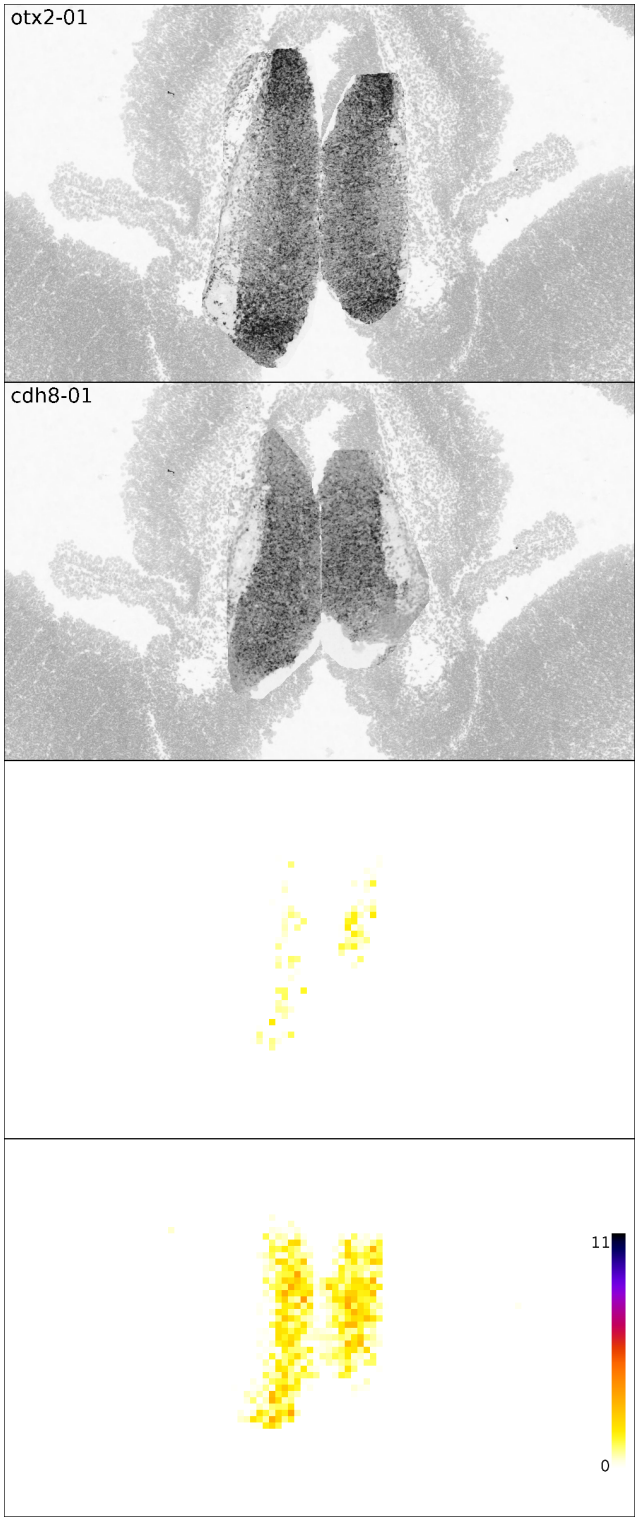


Figure 3.15: Otx2-Cdh8. Rostral end data.

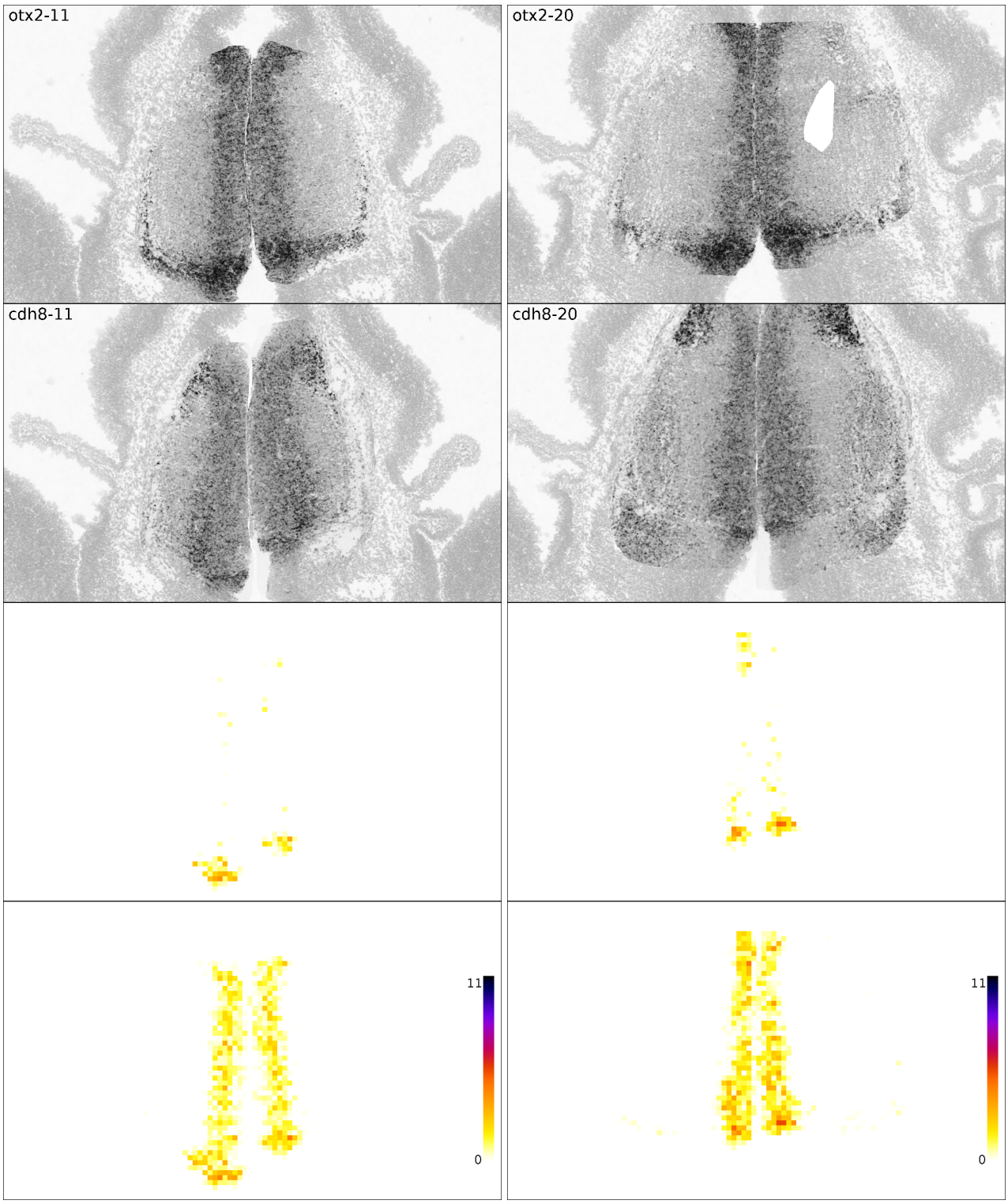


Figure 3.16: Otx2-Cdh8. Rostral data.

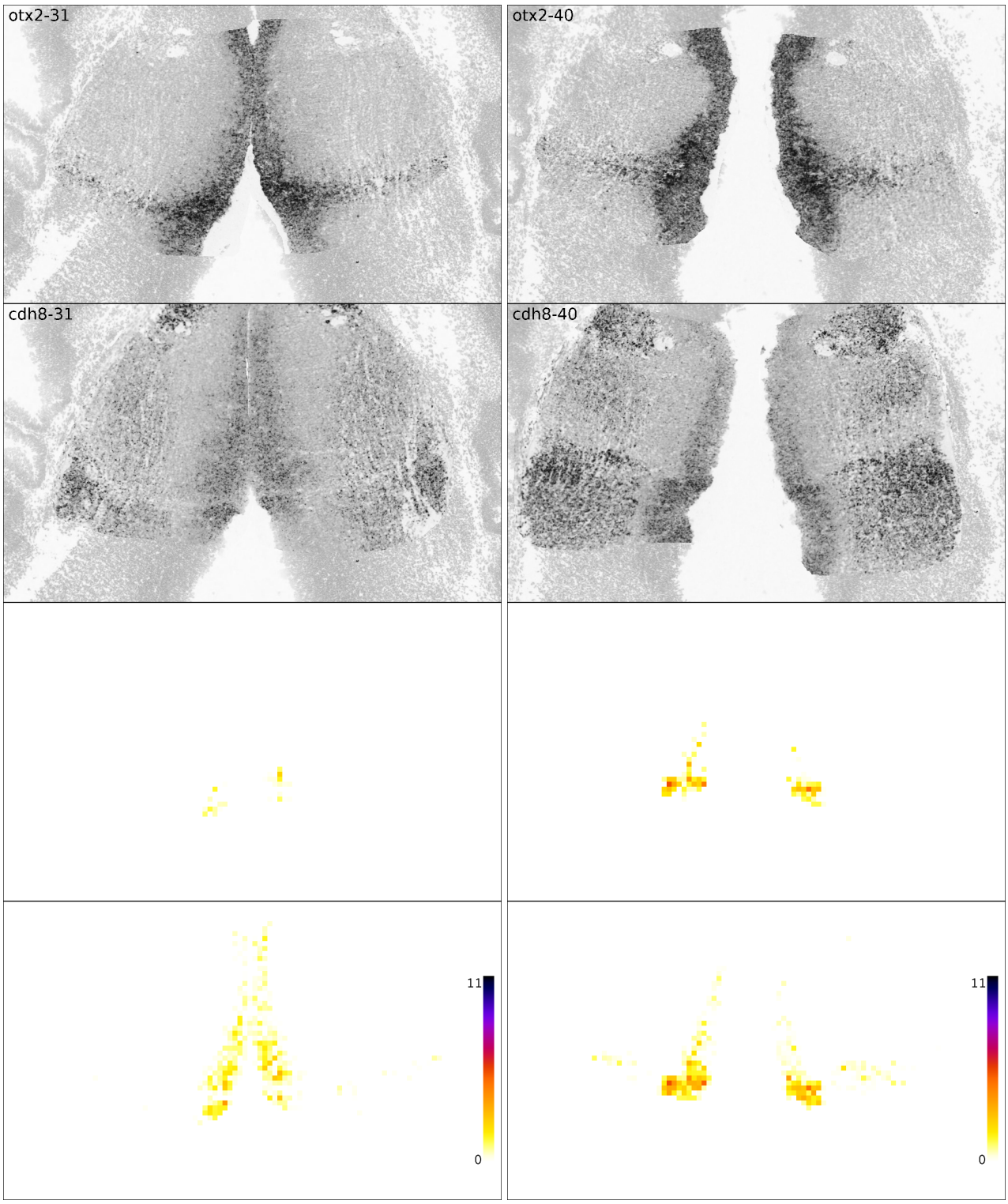


Figure 3.17: Otx2-Cdh8. Caudal data.

Class A

Ngn2-Cdh8

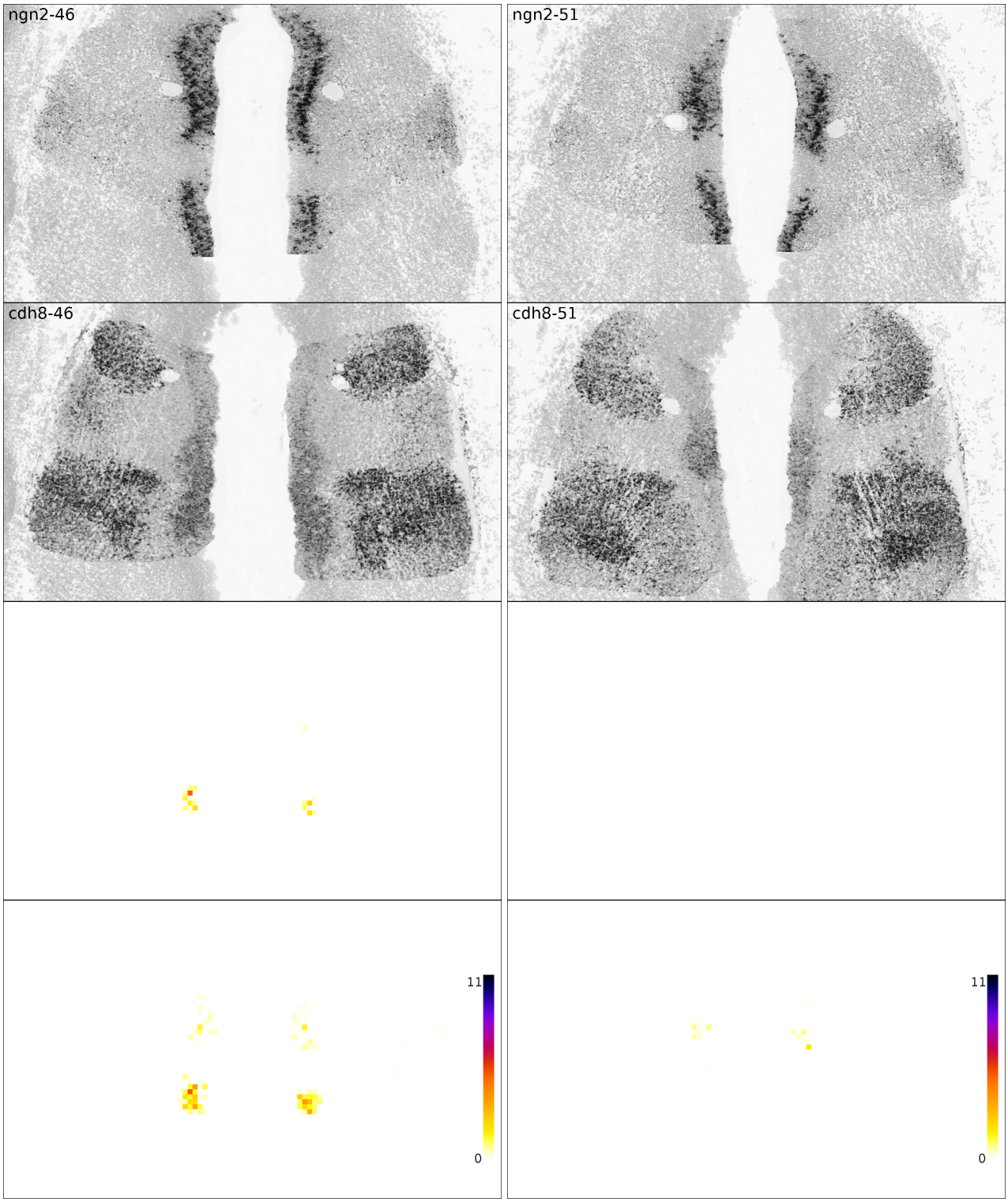


Figure 3.18: Ngn2-Cdh8. Caudal end data.

Gbx2-EphA4

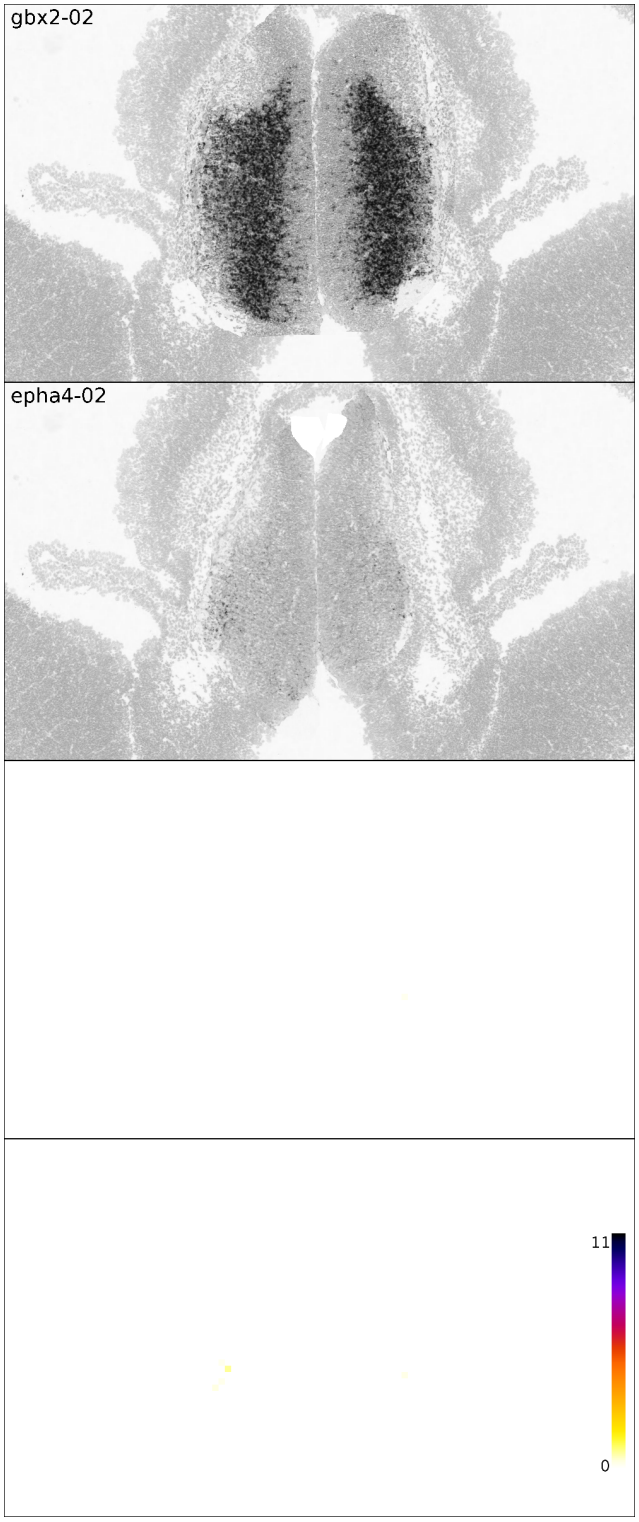


Figure 3.19: Gbx2-EphA4. Rostral end data.

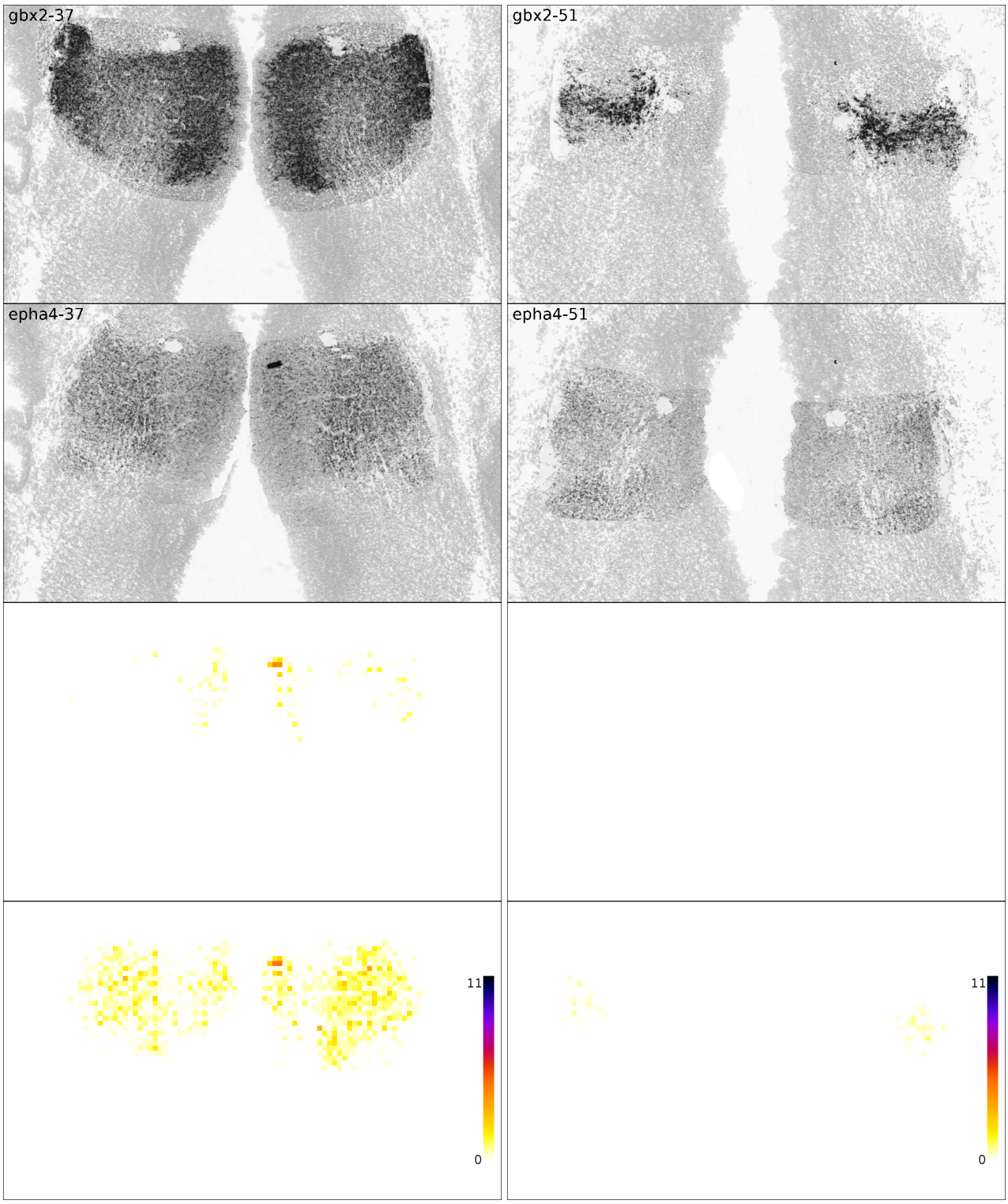


Figure 3.20: Gbx2-EphA4. Caudal (left) and caudal end (right) data.

Gbx2-Ngn2

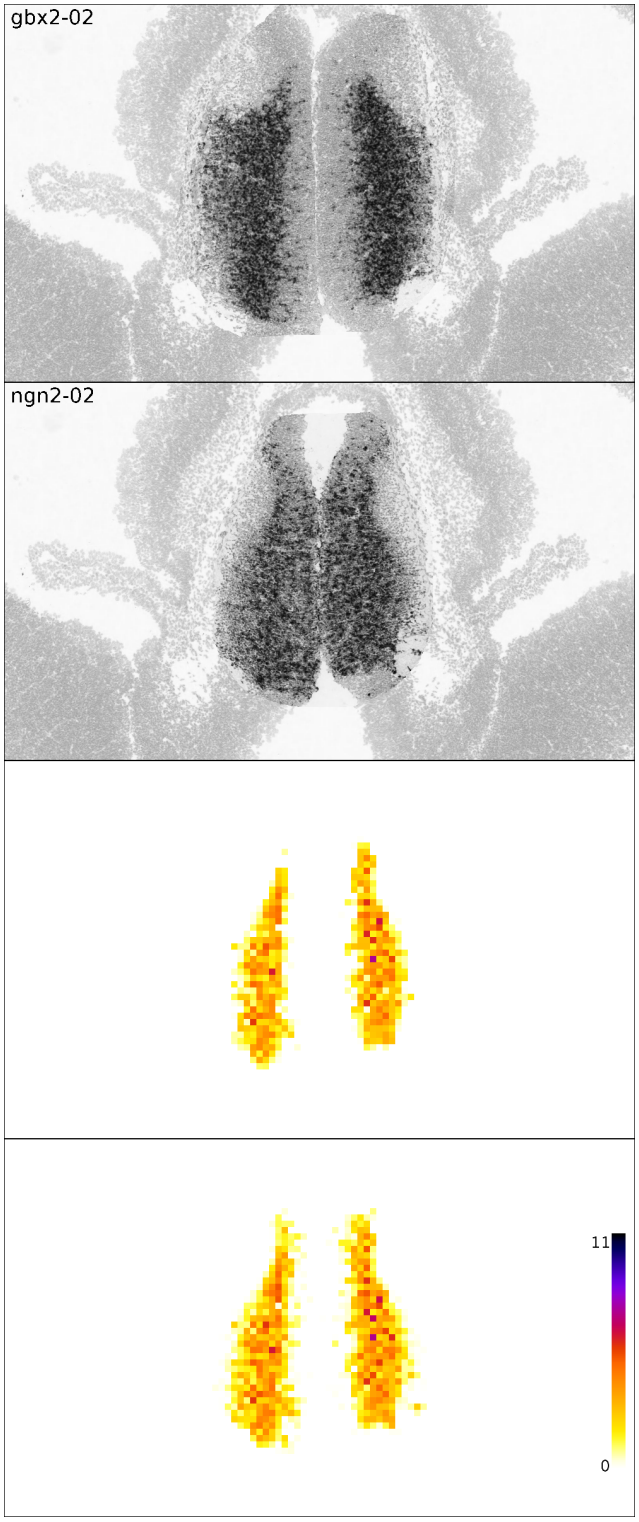


Figure 3.21: Gbx2-Ngn2. Rostral end data.

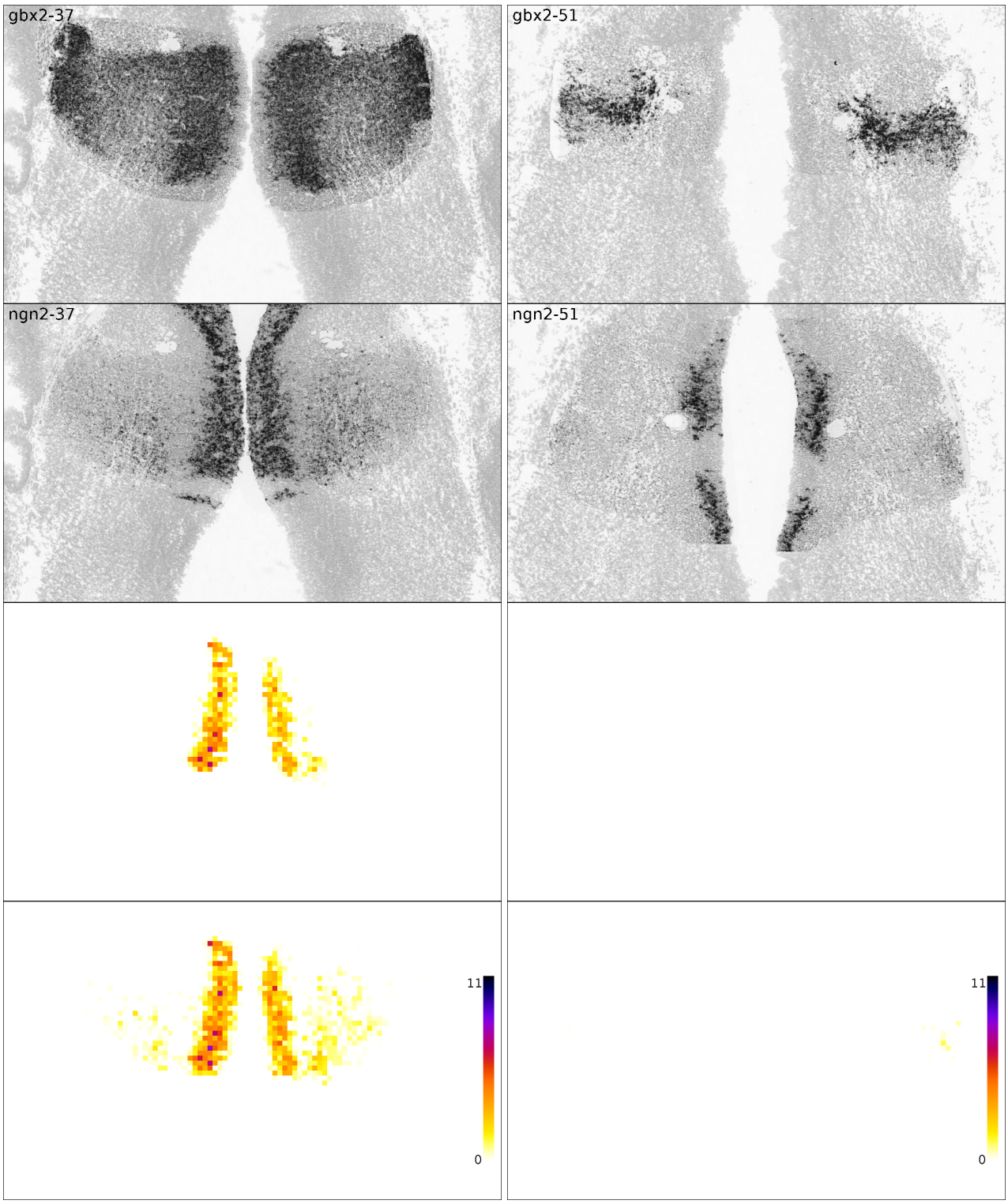


Figure 3.22: Gbx2-Ngn2. Caudal (left) and caudal end (right) data.

Ngn2-EphA4

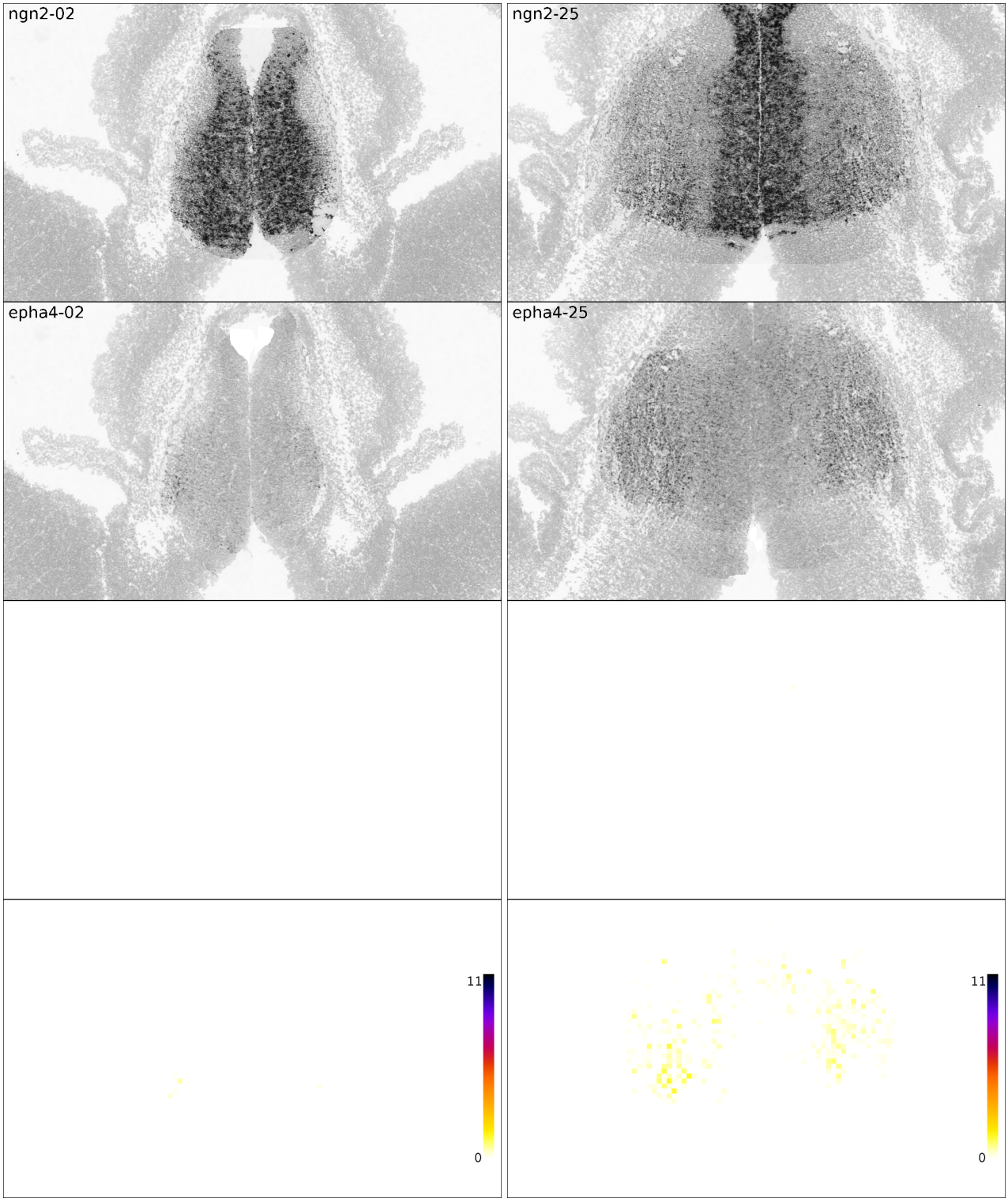


Figure 3.23: Ngn2-EphA4. Rostral end (left) and rostral (right) data.

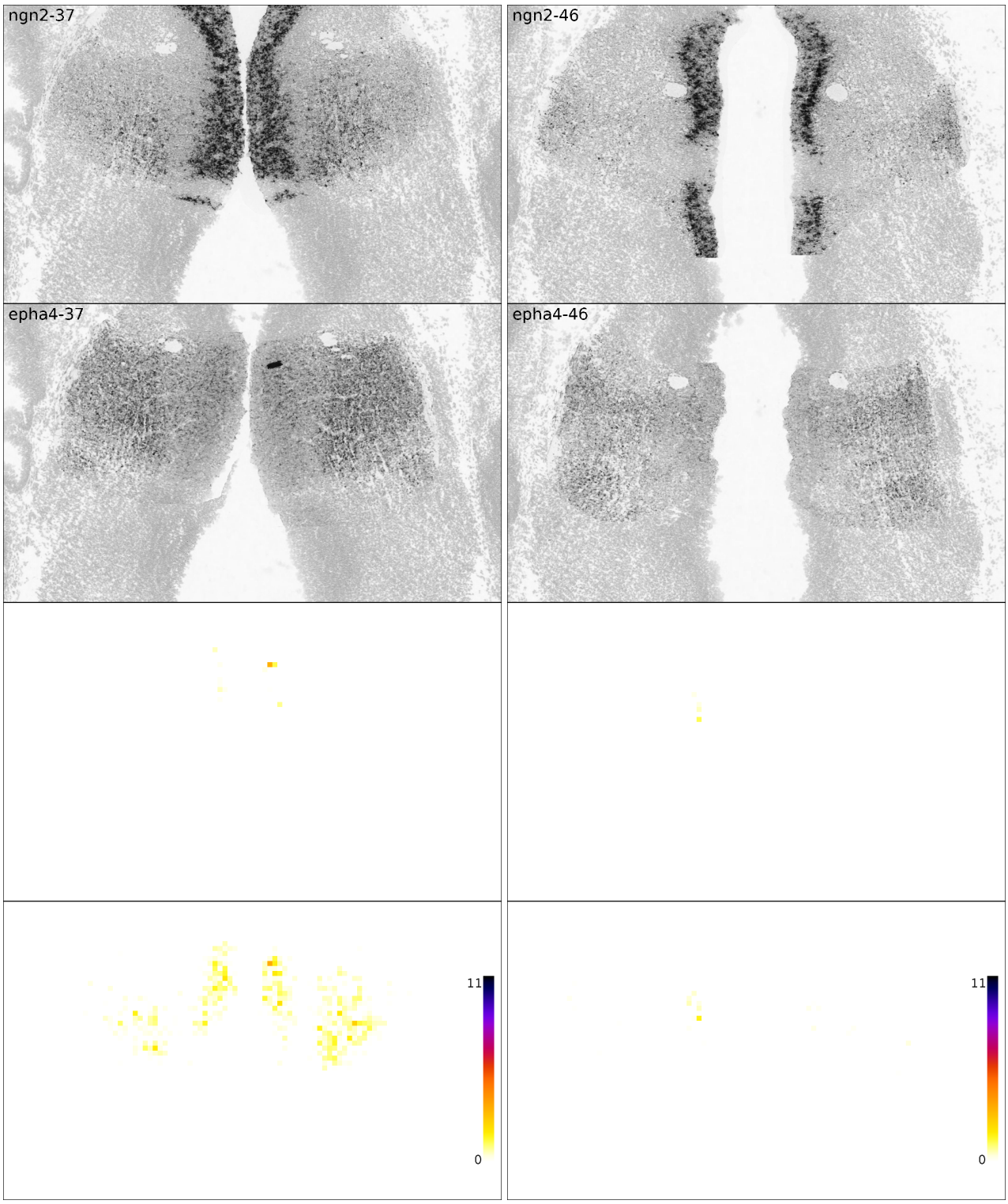


Figure 3.24: Ngn2-EphA4. Caudal (left) and caudal end (right) data.

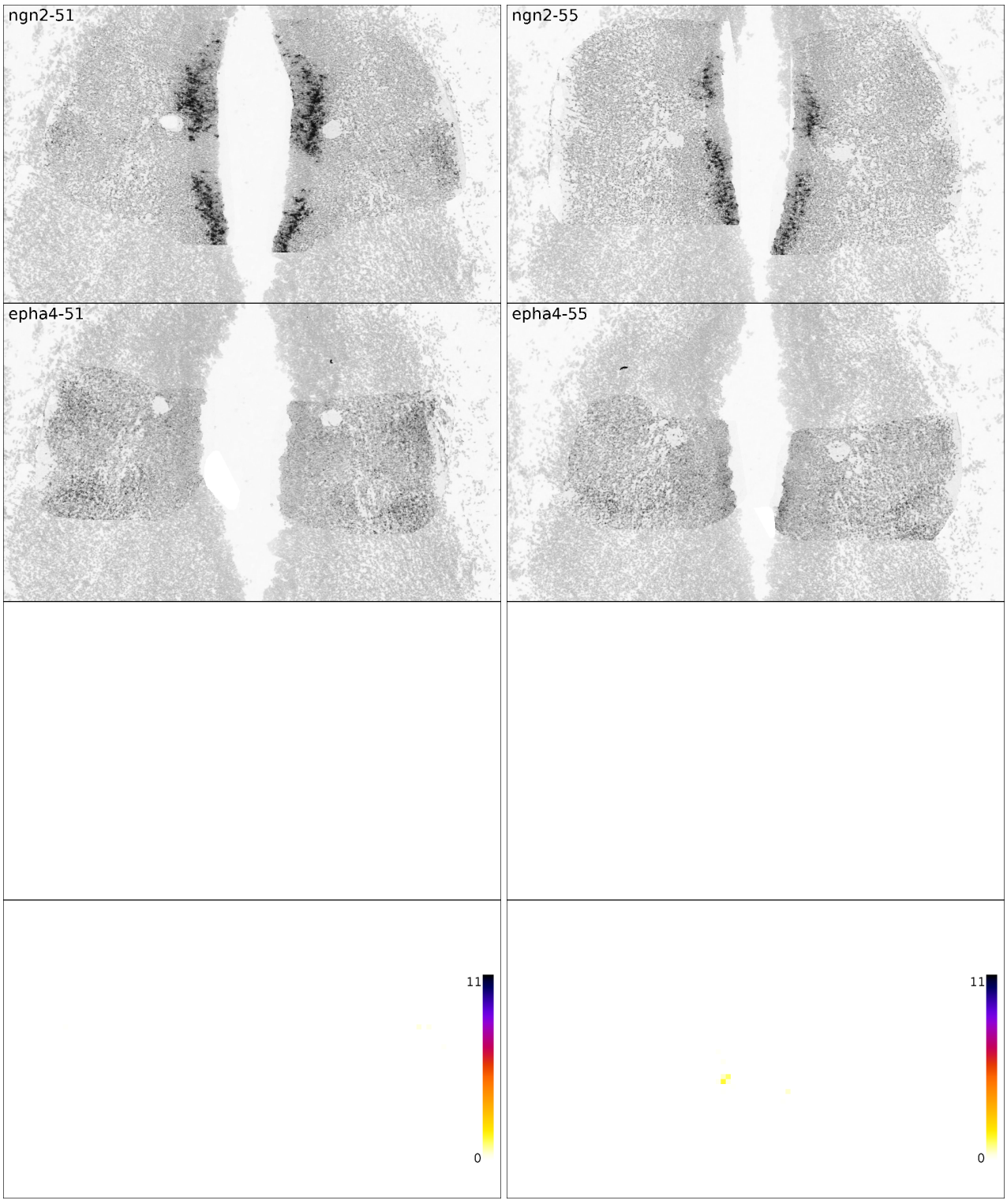


Figure 3.25: Ngn2-EphA4. Caudal end data.

Class B

Olig2-EphA4

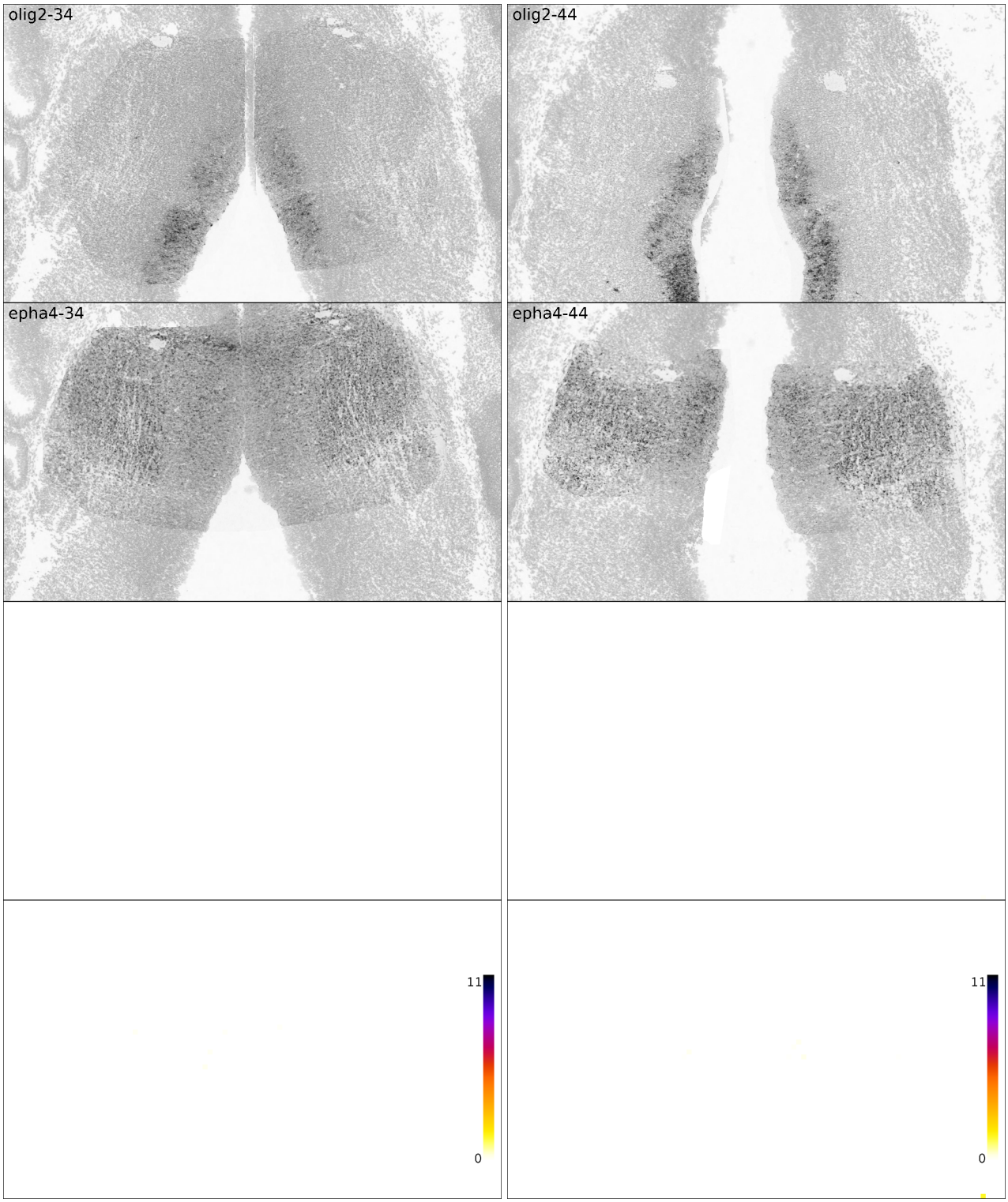


Figure 3.26: Olig2-EphA4. Caudal data.

Gbx2-Olig2

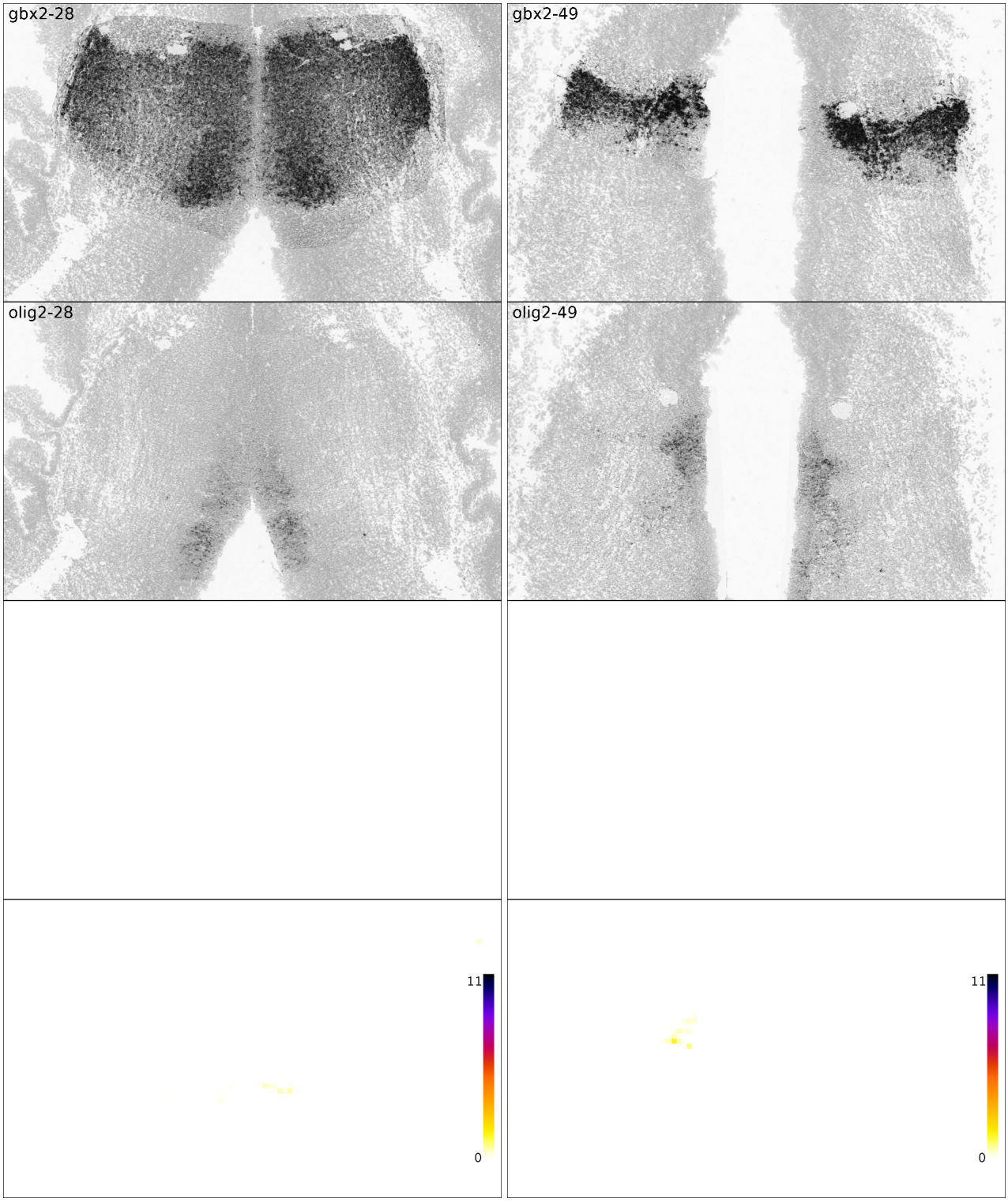


Figure 3.27: Gbx2-Olig2. Caudal (left) and caudal end (right) data.

Olig2-Cdh8

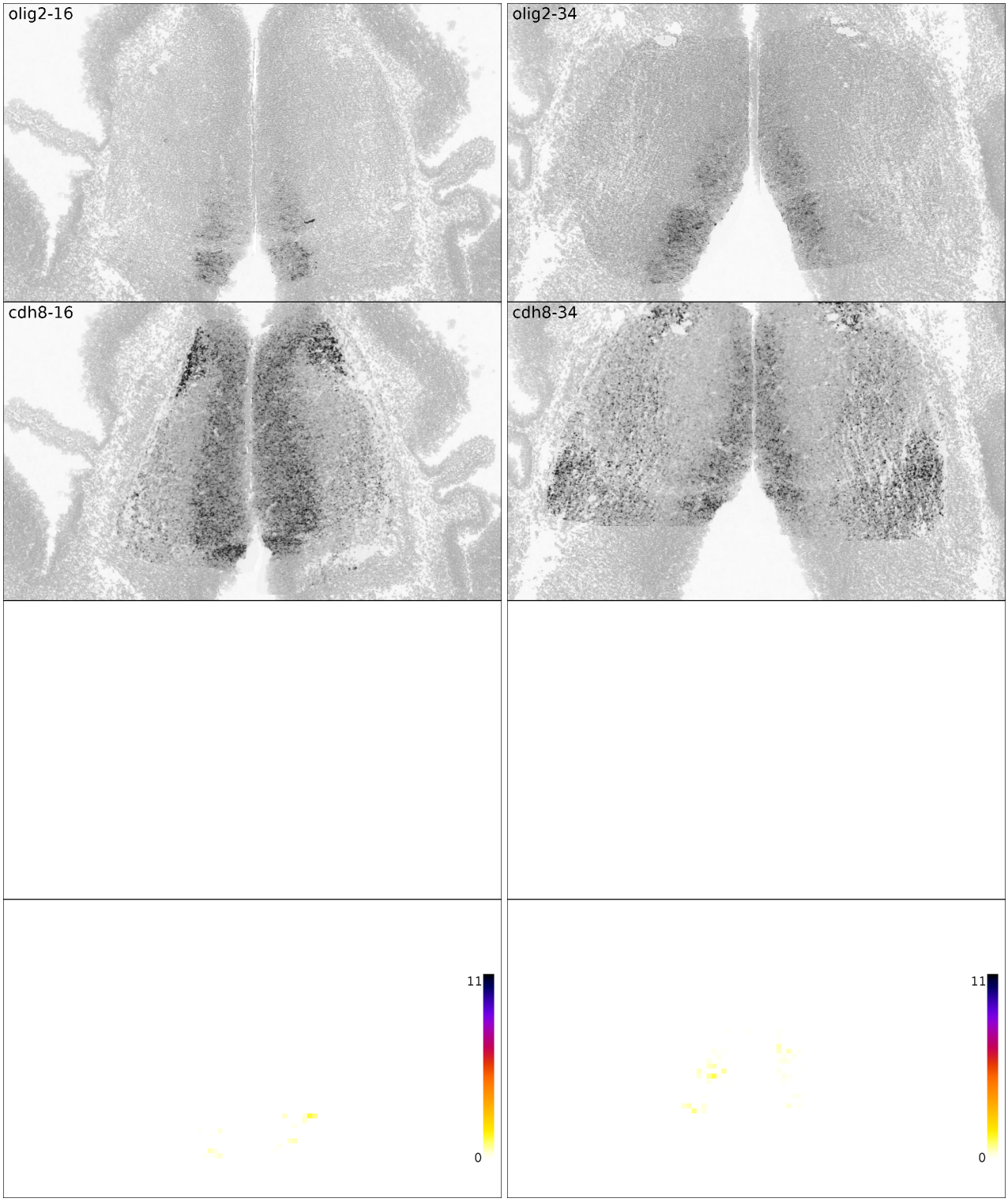


Figure 3.28: Olig2-Cdh8. Rostral (left) and caudal (right) data.

Cdh8-EphA4

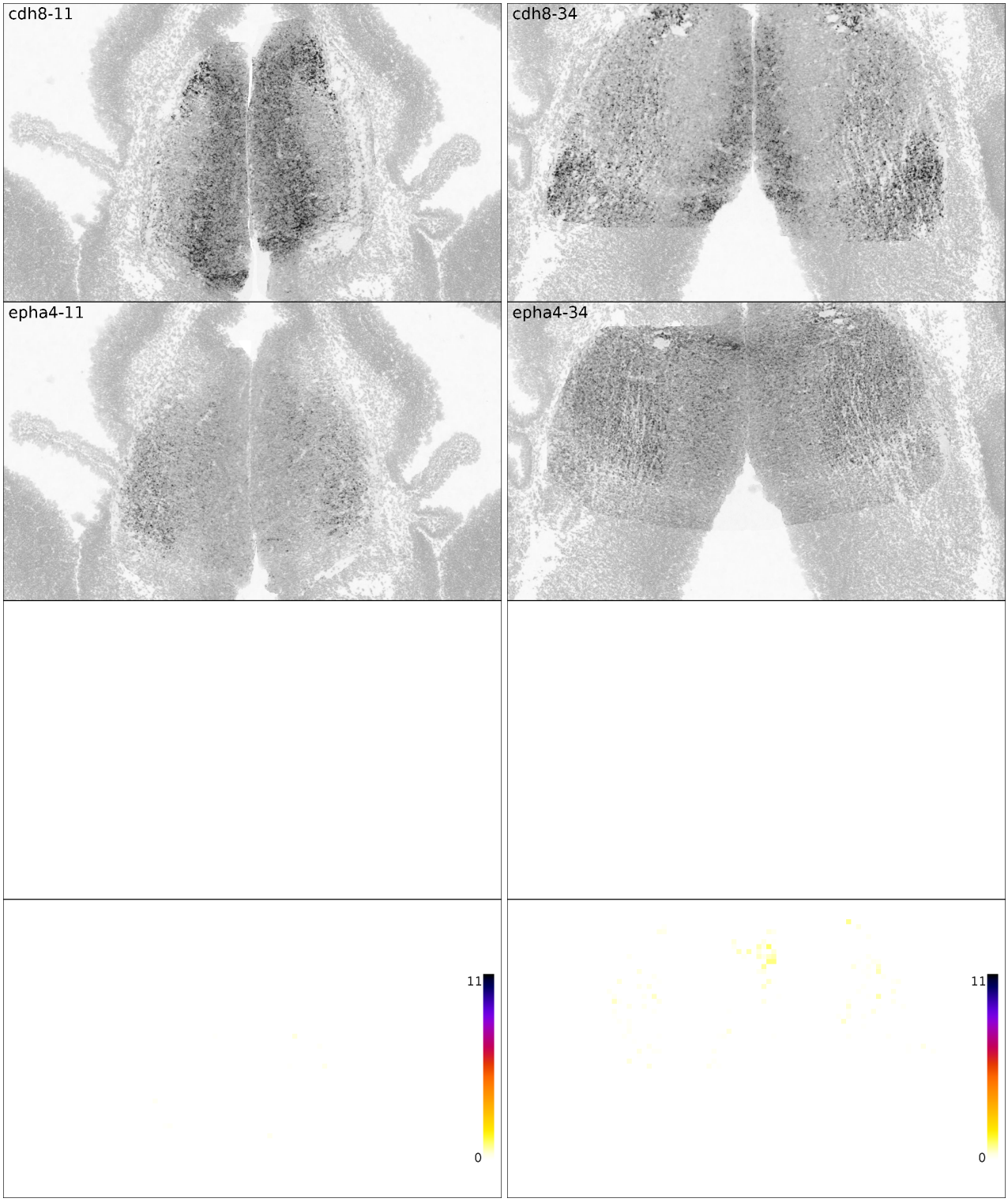


Figure 3.29: Cdh8-EphA4. Rostral (left) and caudal (right) data.

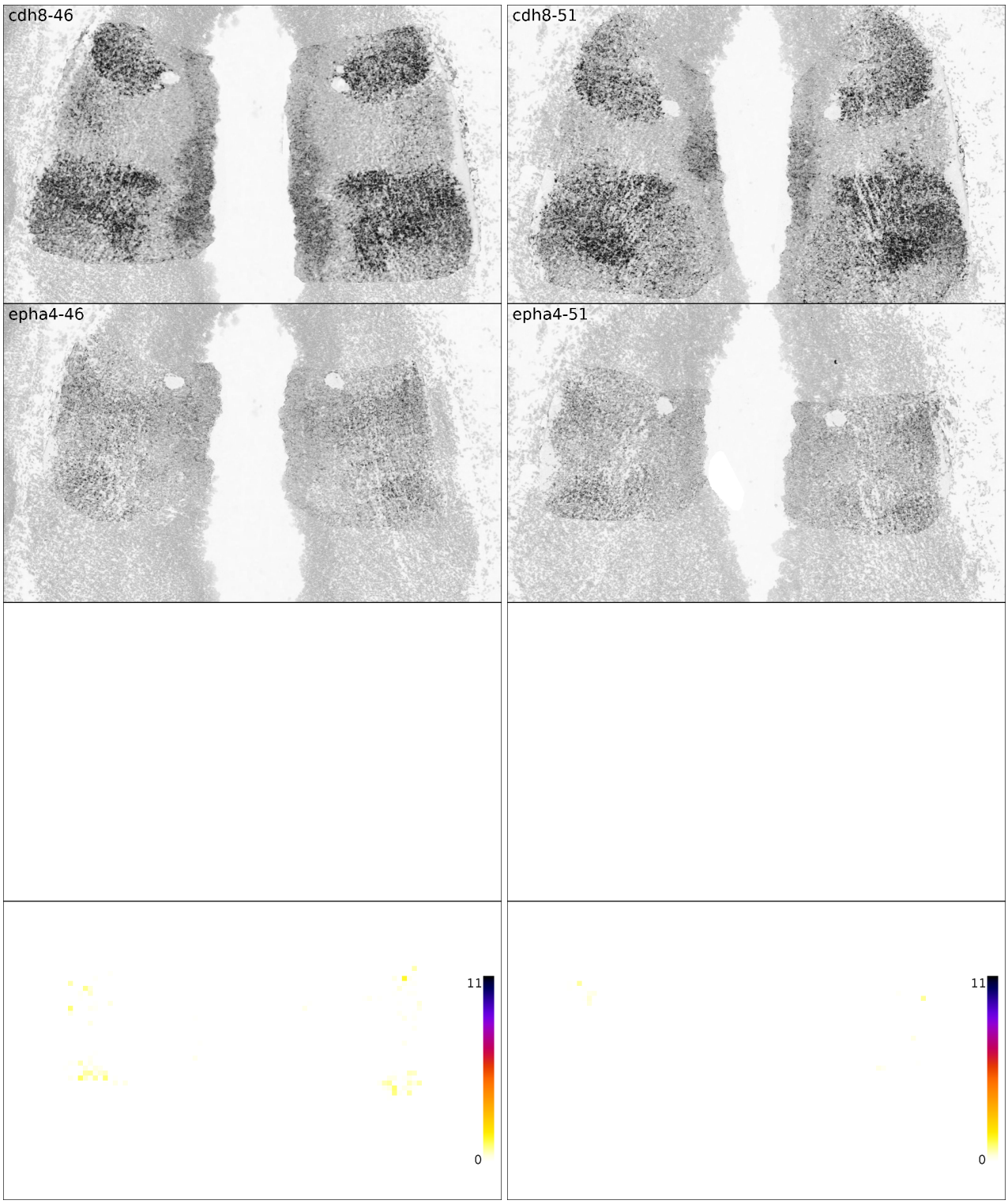


Figure 3.30: Cdh8-EphA4. Caudal end data.

Otx2-EphA4

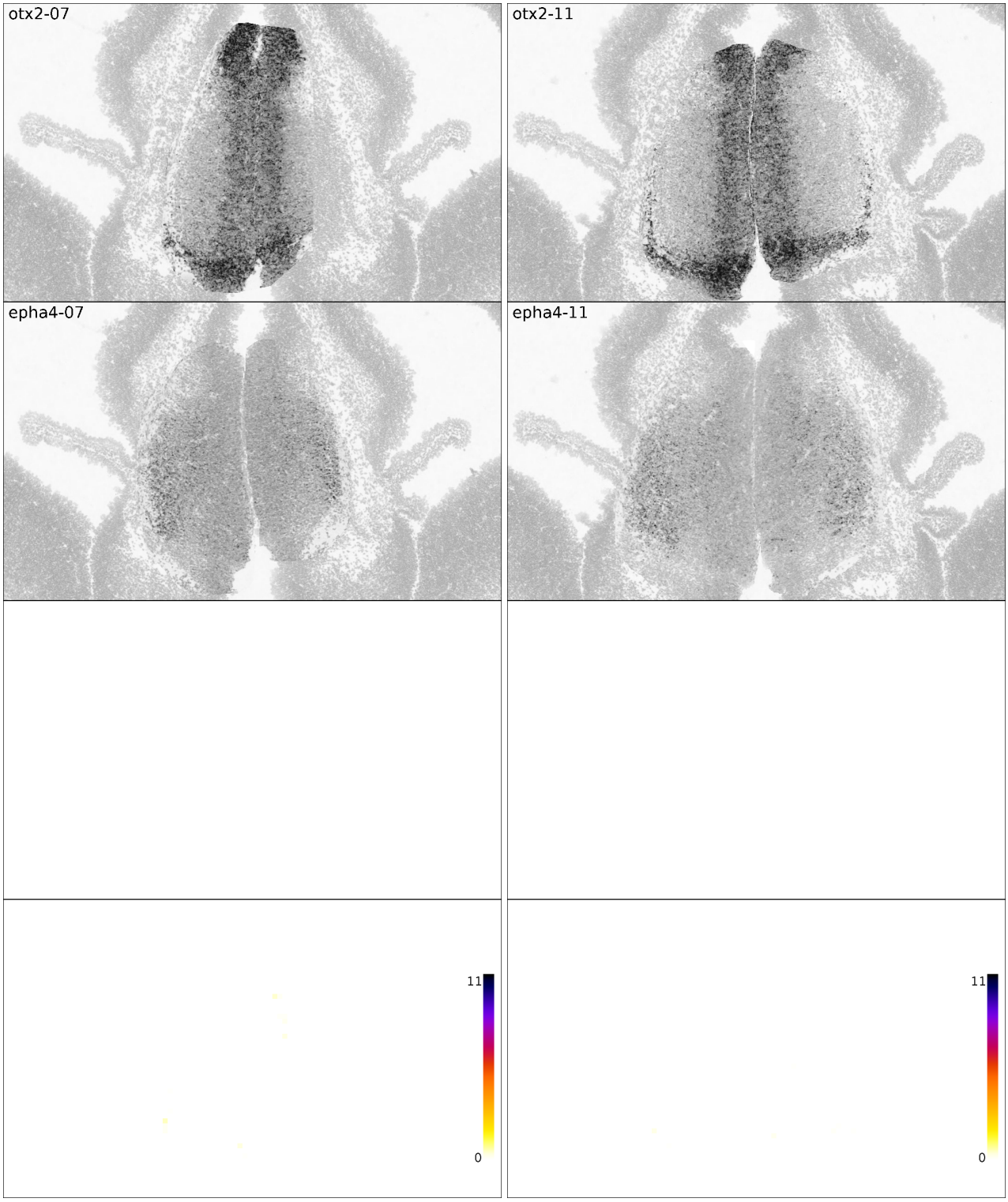


Figure 3.31: Otx2-EphA4. Rostral end (left) and rostral (right) data.

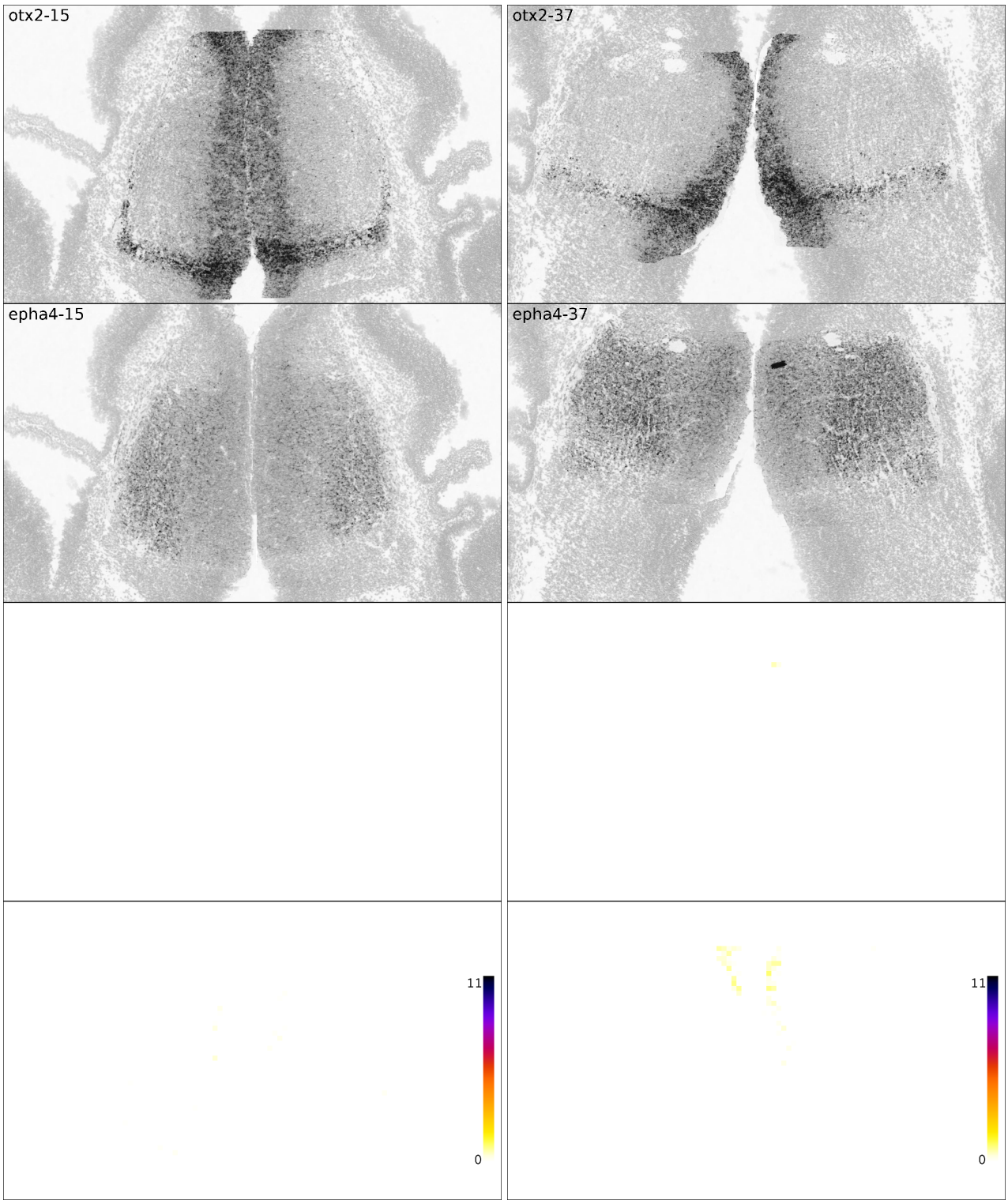


Figure 3.32: Otx2-EphA4. Rostral (left) and caudal (right) data.

3.4 Result analysis

In the previous chapter a procedure to combine gene expression information for two genes and the associated cell density was first presented and then applied to the chosen genes in the E13.5 thalamus. Each pair of genes was categorised according to their co-expression relationship, and a coarse description of the location and the kind of co-expression was provided.

This section brings the various results together, making it possible to compare different co-expression patterns in a meaningful approach. Conclusions drawn from the aggregated data will then be used to provide insights on how the thalamus might develop at this age.

3.4.1 Aggregation

Table 3.6, based on table 3.5, elaborates further on the description of the co-expression between gene pairs. This description is schematic to make sure the table fits in a single page and that different cases are as easy to compare as possible.

A variety of regions and shapes are used (such as cluster, stripe, scatter) and the anatomical location is specified. For each of them, a kind of co-expression is assigned: existent, potential, and potential but unrealistic.

Building on table 3.6, the different co-expression patterns are presented on the corresponding simplified diagrams of the thalamic location in the rostro-caudal axis. Figures 3.33, 3.34, 3.35 and 3.36 contain the different gene pair patterns for the rostral end, rostral, caudal and caudal end areas of the thalamus respectively. Three gray levels are used to indicate the three co-expression levels described previously, and gene pairs that share a pattern are grouped together.

Additionally, an annotated coronal section of the corresponding thalamic area from the Allen [Adult] Brain Atlas is included under each figure to allow quick comparison between the two. Direct connections between the two cannot be made because brains at E13.5 and P56 are substantially different, and some areas of the embryonic thalamus will disappear in the adult, such as the proliferative zone in the midline. Groups of nuclei can still be linked to areas to appreciate how co-expression patterns might relate to future nuclei given that they both can be described in terms of basic positioning via consistent sectioning axes (rostro-caudal and dorso-ventral) that preserve the space distribution throughout these stages. This approach, seen in ref. [21], is used here too to define potential relationships.

	Rostral end	Rostral	Caudal	Caudal end
Section#	1-10	11-27	28-45	46-55
Class B				
Olig2			xx	
EphA4			VM cl	
Gbx2			x	x
Olig2			VM cl	
Olig2		x	x	
Cdh8		VM cl		
Cdh8		x	x	xx
EphA4		Msc	thick M+Lst	DM cl
Otx2	x	xx	x	
EphA4	*thin M+Vst*	*Msc*	top Mst	
Class A				
Ngn2				xx
Cdh8				Mst, *Lsc*, **VM S cl**
Gbx2	x		x	x
EphA4	*VMsc*		**DM+Dst**, broadTh	DLcl
Gbx2	x		x	x
Ngn2	**B ctr DVst**		**Mst**, Vsc	*Lcl*
Ngn2	x	x	x	xxx
EphA4	VM S cl	ctr sc	**DM S st**, ctr sc	*Lsc*, Mst
Class A+				
Ngn2			x	
Otx2			**Mst (Vwider)**	
Olig2		x		x
Otx2		**VMcl**		**ctr Mcl**
Gbx2	x		xx	
Otx2	**B ctr DVst**		**Mst**	
Gbx2	xx		x	x
Cdh8	**B ctr DVst (Vwider)**		**Mst**, Lsc	**Dst**
Otx2	x	xx	xx	
Cdh8	**ctr DVst**, wider Vst	**VMcl, *Mst,*** Vst	**VMcl (Vgrow)**, M+Vst	

Table 3.6: Co-expression description for the gene pairs in a schematic form.

First two rows: subdivision name and section range. For each gene pair, the first row indicates the section count per division and the second one contains the description. M/L/D/V: Medial/Lateral/Dorsal/Ventral; S/B: Small/Big; cl: cluster; sc: scatter; st: stripe; ctr: center/central. Emphasis used: *italics* - very faint possible coexpression; regular - possible coexpression; **bold** - definite coexpression.

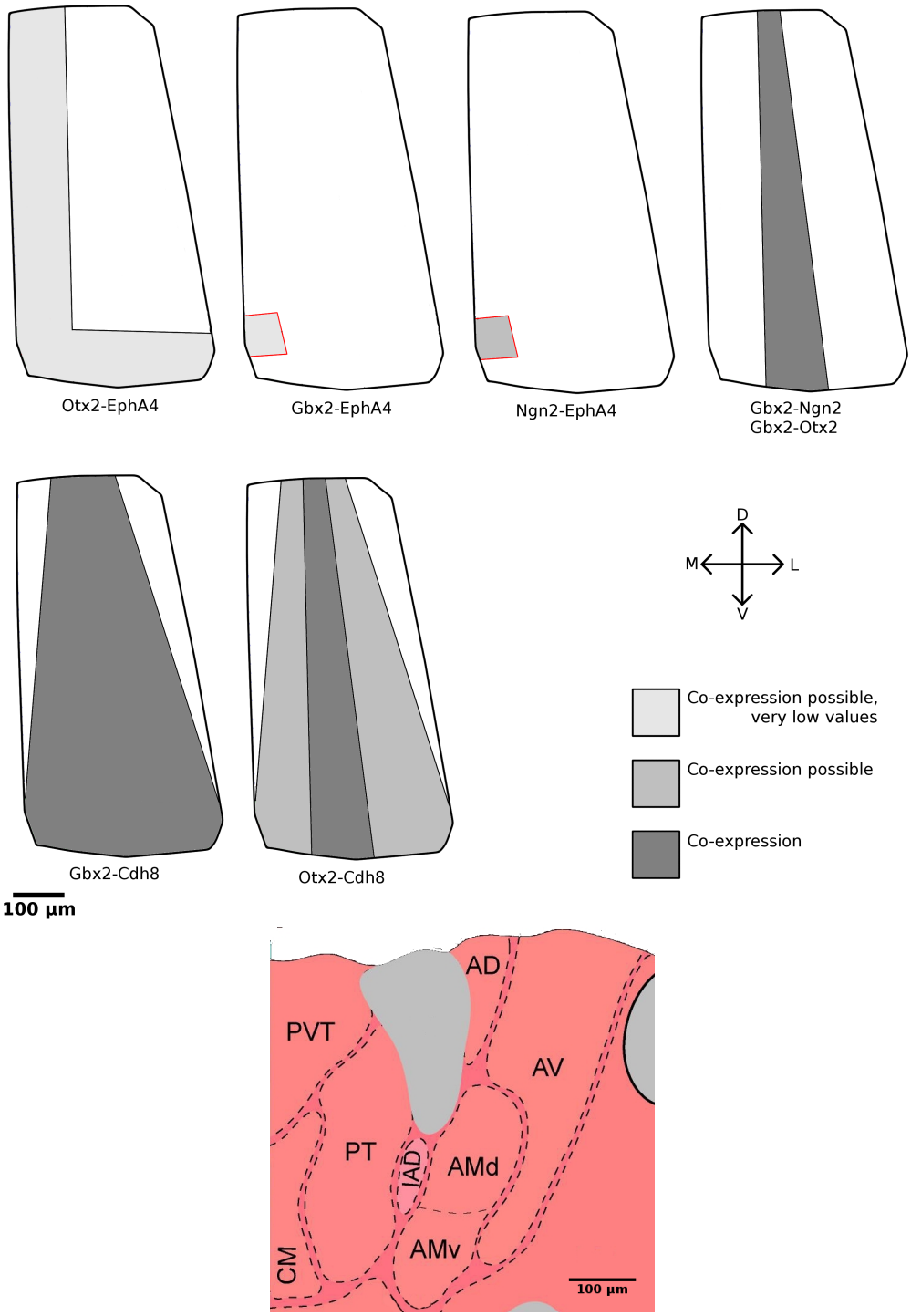
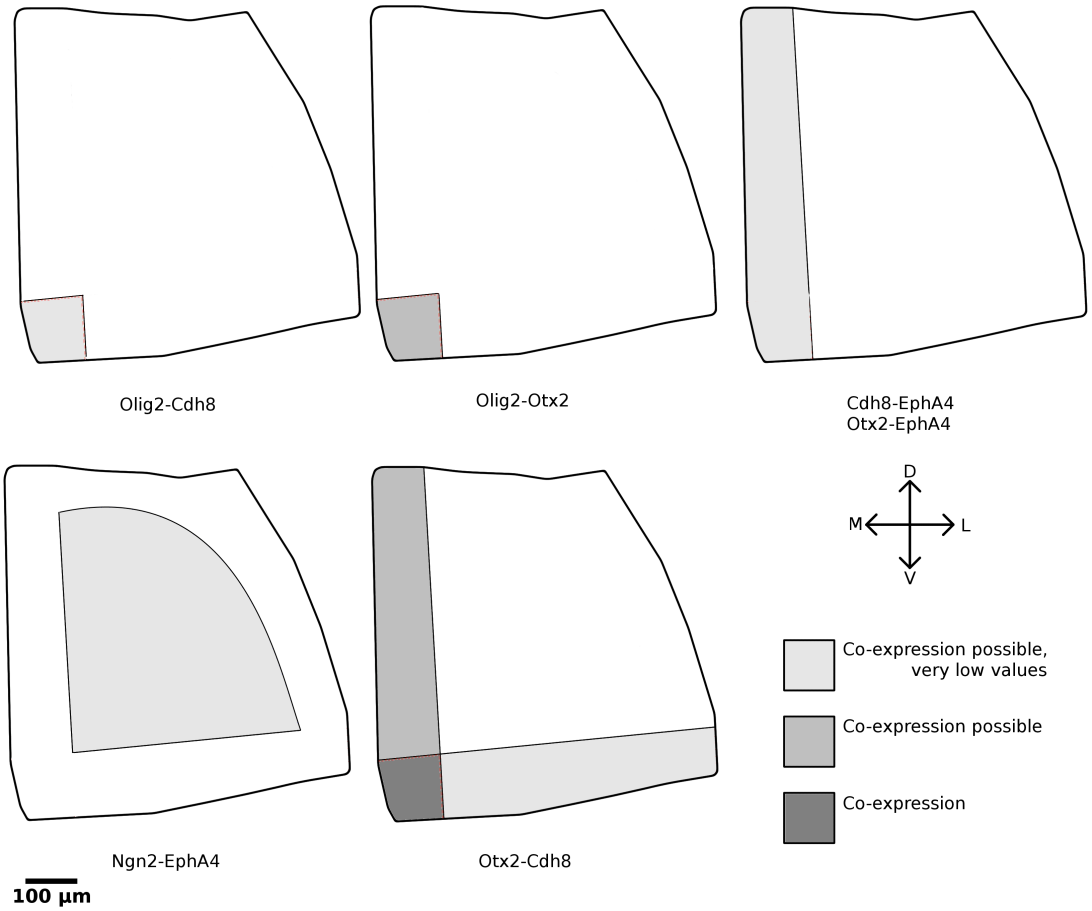


Figure 3.33: Gene pair co-expression patterns at the rostral end.
Top: co-expression patterns. Most of them are based around the center of the thalamus. The medial and ventral borders and specially the ventro-medial area also show some co-expression.
Bottom: corresponding adult area, annotated, from the Allen Mouse Brain Atlas (AMBA).



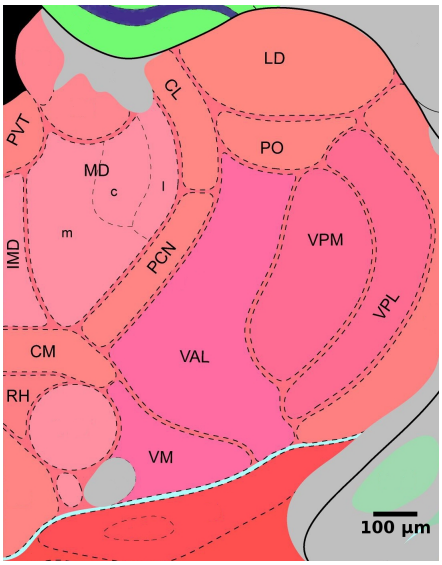


Figure 3.34: Gene pair rostral co-expression patterns.
Top: co-expression patterns. Co-expression is present basically on the medial and ventral borders and the dorso-medial area, except for the Ngn2-EphA4 case where there is general central scatter.
Bottom: corresponding adult area, annotated, from the AMBA.

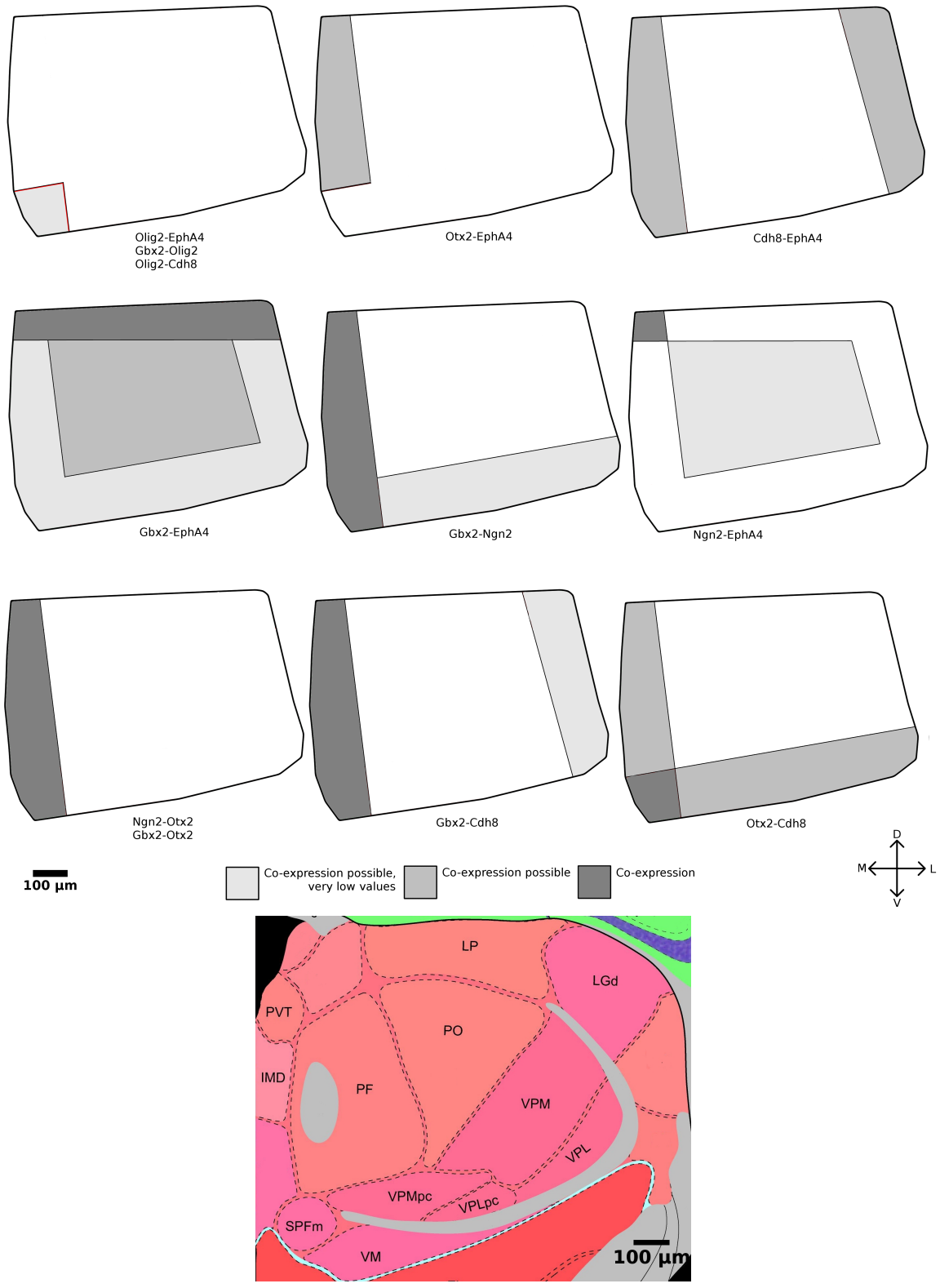


Figure 3.35: Gene pair caudal co-expression patterns.
Top: co-expression patterns. Co-expression exists everywhere at this location, with the medial border and its dorsal and ventral ends concentrating most of the cases.
Bottom: corresponding adult area, annotated, from the AMBA.

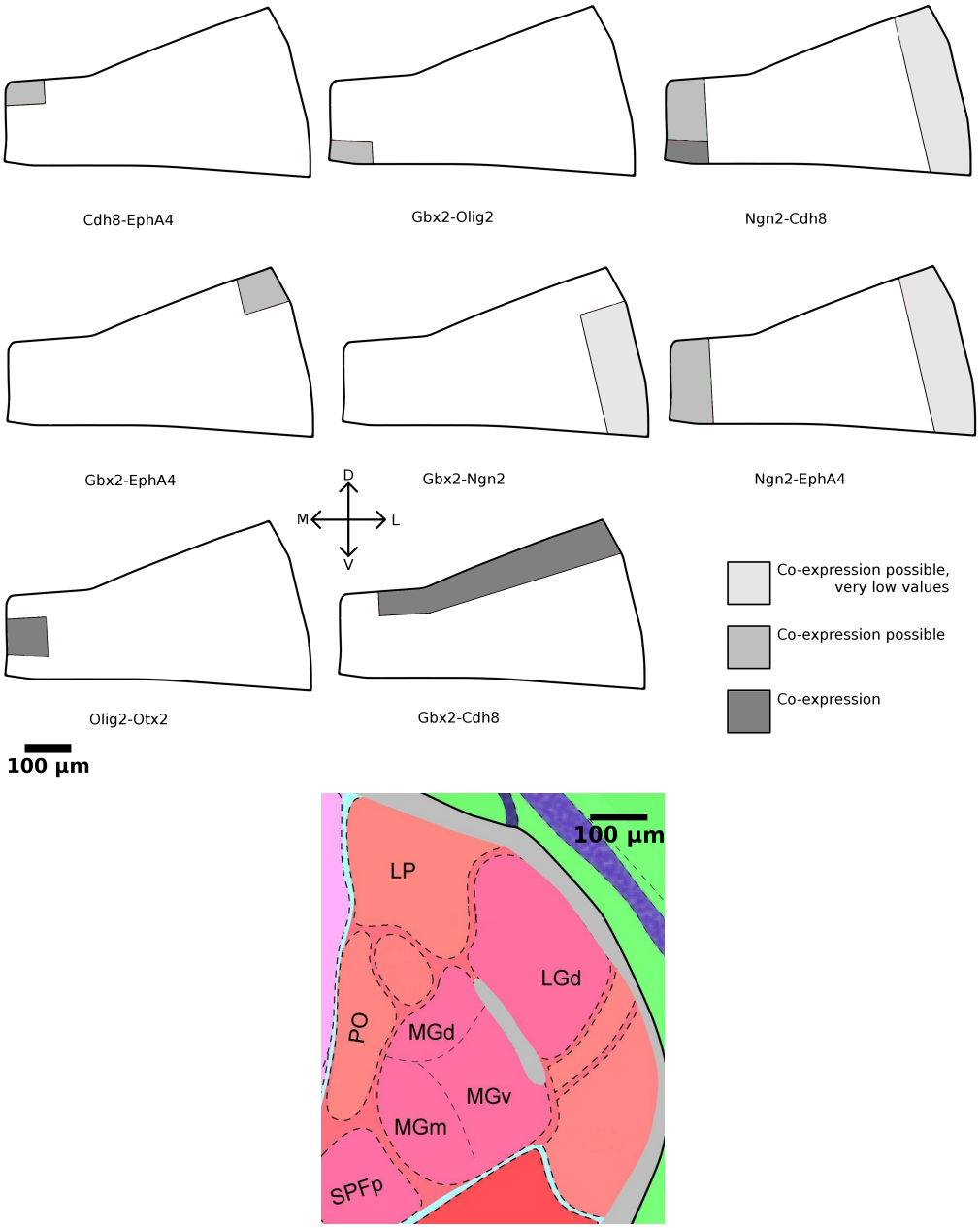


Figure 3.36: Gene pair co-expression patterns at the caudal end.
Top: co-expression patterns. Co-expression patterns are much more localised and concentrate around the medial, lateral and dorsal borders.
Bottom: corresponding adult area, annotated, from the AMBA.

3.4.2 Description

The **rostral end** has the biggest change in anatomy, making it more difficult to compare both figures. Gbx2 and Cdh8 are co-expressed by most of the cells. The central vertical stripe defined by Gbx2-Ngn2 and Gbx2-Otx2 closely match the AD and AM nuclei. Otx2-Cdh8 are also co-expressed in that same stripe but potentially most of the thalamus. Otx2-EphA4 cells cover the midline and ventral areas, matching directly the PVT, CM, and ventral parts of the PV, AM and AV. Additionally, the centromedial blob defined by Gbx2-EphA4 and Ngn2-EphA4 also belongs to the CM.

At the **rostral** part, embryonic data shows the a layout that reflects closely the adult. Otx2-EphA4 and now also Cdh8-EphA4 and Otx2-Cdh8 cover the midline area, matching the PVT, MD, CM and RH. Otx2-Cdh8 additionally spread through the ventral border, covering VM, VPL and VAL. The centromedial blob specified by Olig2-Cdh8 and Olig2-Otx2 covers RH and CM. Finally, Ngn2-EphA4 cells spread over the central thalamic space, matching MD, PCN, PO, VPM and parts of LD, CL and VAL.

The **caudal** part is organised similarly to the previous one. Gbx2-EphA4 cover all the thalamic space, and are specially strong dorsally. The midline is covered by most of the gene pairs, and it is closely linked to PVT, MD, PF, SPF and VM. The lateral border, corresponding to LP, LGd and part of VPL, is made of Cdh8-EphA4 and Gbx2-Cdh8 cells. Otx2-Cdh8, Gbx2-Ngn2 and Gbx2-EphA4 do also cover the ventral band, associated with SPF, VM and VPL. The central space is only influenced by the gene pairs Gbx2-EphA4 and Ngn2-EphA4 and covers PF, PO, VPM and part of VPL.

At the **caudal end**, the adult layout looks very similar, if only rotated. Gbx2-Cdh8 co-expressing cells are located around the dorsal end, linked to PO and LP. Ngn2-Cdh8 and Ngn2-EphA4 are located at both medial and caudal ends (Gbx2-Ngn2 only lateral), covering SPF, LP and LGd. Cdh8-EphA4, Gbx2-Olig2 and Olig2-Otx2 are located in different clusters on the medial end, but they all are closely related to the SPF. Finally, Gbx2-EphA4 are located on a dorsolateral patch that fits on the LP nucleus.

Table 3.7 aggregates the described matches between the co-expression patterns of gene pairs and nuclei. Pairs that are consistently co-expressed on a prospective nucleus in more than one rostro-ventral subdivision are shown in bold.

Group	Nucleus	Rostral End	Rostral	Caudal	Caudal End
Anterior	AD	GN,GT,TC,GC	-	-	-
	AM	GN,GT,TC,GC	-	-	-
Hypoth	PVT	**TE**,**GC**	**TE**,**CE**,**TC**	**TE**,**CE**,**TC**,**GC**,GN,NE,NT,GT,GE	-
Intra-Laminar	PCN	-	NE	-	-
	CM	**TE**,GE,NE,GC	**TE**,TC,CE	-	-
	RH	-	OC,OT,OE,CE,TC	-	-
	CL	-	NE	-	-
	SPF	-	-	**CE**,**GO**,OE,OC,GN,NT,GT,GC,TC,GE	**CE**,**GO**,NC,NE,OT
	PF	-	-	GE,NE	-
LGd	LGd	-	-	CE,GC	NC,GN,NE
LP	LP	-	-	**GE**	**GE**,NC,NE,GC
Medial	MD	-	**CE**,**TE**,**TC**,NE	**CE**,**TE**,**TC**,GN,NT,GT,GC,GE	-
MGc	MG	-	-	-	-
PO	PO	-	**NE**	**NE**,GE	GC
Ventral	VM	-	**TC**	**TC**,GN	-
	VAL	-	TC,NE	-	-
	VPL	-	**TC**	**TC**,GN	-
	VPM	-	**NE**	**NE**,GE	-

Table 3.7: Co-expression patterns and prospective nuclei within each of the groups.
For each thalamic division in the rostro-caudal axis, the gene pairs whose co-expression patterns cover the are matching the future position of the nucleus of each row are shown. Gene pairs that appear on more than one division are in bold to indicate a general presence in the thalamus.
Genes: T Otx2, E EphrinA4, G Gbx2, C Cdh8, N Ngn2, O Olig2.

3.5 Discussion

Functions to assess the minimum and maximum amount of gene co-expression given two comparable *in situ* sections have been defined. By combining them with the cell density measurements of the reference dataset calculated with the Hessian-based method (also shown in the previous chapter), cell numbers are obtained per each sampled tile.

Imaged sections from genes used in this project have been integrated into the space of the reference dataset, and their level of expression has been sampled using the tool described in the previous chapter. Each pair of genes has had its minimum and maximum number of co-expressing cells per tile measured, and it has been assigned a class type based on the kind of co-expression relationship they share: no co-expression, co-expression possible, necessary co-expression. Finally, the resulting figures have been presented grouped by class and ordered by quantity and quality of contained results.

3.5.1 Result validation

The most important future work of this project would be the validation or confirmation of the results obtained about co-expression of genes.

The best way to achieve that would be using double immunohistochemistry or *in situ* hybridisation

protocols. IHC, the easiest method if it works, would require having working antibodies for the desired genes (that is not always the case), considerable fine tuning, and luck. For ISH a more demanding protocol is required, but the required probes do exist, since the Allen Institute used or created them.

3.5.2 Classification criteria

The classification of gene pair datasets described in the previous section could be done following other criteria. Depending on the kind of tissue where the procedure is applied or quality of data, perhaps using the composite of maximum values in both minimum and maximum double labeling is not the most useful way to separate gene relationships.

It would be possible, for instance, to take into account the values above zero for each section and use their distribution as means to discard gene pairs that only belong to a higher class because of noise or problems arising from the variation in slicing. The number of sections per pair could also be incorporated as a measure of reliability of the data. It is also likely that the combination of these and other criteria could be adapted to the specific needs of a given project to result in the best separation of relationships.

3.5.3 Comparison to previous study

Table 1.2 in the first chapter contained the expression classification of several genes within the early postnatal mouse thalamus done in ref. [21]. In this part of their study, they visually assessed whether each of four genes was expressed, weakly expressed or had non-detectable expression in each of the thalamic nuclei detectable at P2.

There is an overlap of two genes between the work presented in this thesis and the one they did, specifically Gbx2 and Ngn2, so it is interesting to see how the results obtained via such different methods might compare to one another. The main difference lies in the ages used and the method to decide the region of expression. This study uses E13.5 and relates to the adult, while they focused on P2 and compared gene expression with a more accurate assessment of tissue subdivision: cytochrome oxidase, which reveals cell activity and in this case early nuclei.

Ngn2 and Gbx2 are co-expressed in eight of the nuclei listed in table 3.7. Five of them appear in the other study, VP, LGN, AM, MD and VM. Interestingly enough, it is only in one case, AM, that both genes are acknowledged to have expression, while in the other cases one of the two genes is classified as not having detectable expression. This could be attributed to the difference in ages: Ngn2 stops being expressed in early postnatal stage (P4 in the ADMBA shows no Ngn2 expression), and it is likely that Gbx2 might have become more localised by then instead of covering most of the thalamus as seen in the data.

Chapter 4

Discussion

4.1 Discussion

The assembled diagrams for each rostro-caudal segment in the previous chapter reveal the similarities between the different co-expression patterns and provide a clearer visual view of how they might be partitioning the space.

Gene choice

The genes used in this project have proved useful: their co-expression patterns do cover most of the thalamus in each the four key rostro-caudal subdivisions. Results are not perfect though: both rostral areas show poor edge definition, and the caudal end does not reveal much regarding central and ventral parts.

Genes that are expressed widely in the area of interest, such as Gbx2 in the thalamus at E13.5, are of limited use when studying parcellation. The assumption would be that these genes are important for the development of the brain structure but they cover too much to specify any single region. Tissue patterning happens via the interaction of several genes, so genes with more restricted domains should be included in the combinations to study the phenomenon better. The usefulness of these combinations is shown here, where co-expression patterns are linked to prospective nuclei.

Thalamic patterning

E13.5 is a good age to study thalamic development because it is when parcellation starts, but at the same time it is too soon to see many patterns arising. The division of the space based on how co-expression patterns of gene pairs overlap can give an idea of the level of organisation existent at that time, increasing our knowledge of the process.

Throughout the thalamus there appears to be a considerably uniform partitioning of space consisting of edges, a central area and some vertices. The rostral end case presents the exception to this observation, because the space appears divided in vertical bands, but once the three middle ones

are treated as central area the pattern of division becomes similar to the one found in other sections. The ventro-medial cluster, located on top of the ZLI, is revealed as an important piece that is always present. There appears to be a shift in subdivisions containing co-expression while traversing the thalamus caudally: from a ventral-medial bias (rostral end, rostral) the regions cover the whole thalamus (caudal) to finally drift dorsally (caudal end).

Link to nucleation and adult patterning

The progressive gain of importance of the dorso-lateral area caudally can be linked to the growth of the geniculate bodies, and specifically the dorso-lateral geniculate nucleus (dLGN), one of the nuclei that are first visible. The ventro-medial cluster could be seen caudally as the prospective ventro-posterior medial complex (VPM), which also develops soon.

Assuming that the early thalamic parcellation defined as early as E13.5 gives rise to nuclei and that despite obvious changes in size and shape the internal organisation of the thalamus remains similar during late development and until the brain is fully formed, we can link the co-expression patterns described before to the process of nucleation.

Possible interpretations

It is important to note how the cell populations that co-express pairs of genes that are required for thalamic development define such clear regions. Very rudimentary divisions appear in a meaningful way using the medial, dorsal, ventral and lateral borders and especially the vertices where they overlap. These link to the eventual layout of the nuclei, showing that they are most likely the source of the future complexity, a first step towards full parcellation.

Thalamic nuclei start being identifiable from E14.5, so the initial cell clusters that form the protonuclei probably already exist at E13.5. Assuming that the number of initial cell clusters starting parcellation is the same as the final nucleus count in the adult and that protonuclei are similar in size, it would be interesting to estimate the size of the quantity of tissue required to give rise to a nucleus. Table 4.1 contains some basic calculations obtained from data measurements under these assumptions.

Even though the area measurements of the thalamus at E13.5 should be corrected to take into account the proliferative zone -now all the thalamus is included-, there is a striking similarity between estimated protonuclei sizes at the different rostro-caudal levels. These range from 31083 to 52555μm^2, or 176-229μm per side if they were squares (as a reference, the data sampling was done with 20μm sided squares).

The co-expression patterns obtained in this project reveal divisions of the thalamic space that vary considerably in size. It is interesting to see how, in the cases where co-expression is sufficiently strong and located in vertical stripes, their widths range between 100 and 200μm. This could be showing that protonuclei are far from similar in size and shape, which should be expectable, because of underlying

biological complexity or that nucleation is not ready to start at E13.5 . In any case, the study of gene co-expression for a few selected genes grouping them only in pairs has been able to offer an already considerable perspective on early thalamic patterning. The addition of key genes and the inclusion of wider combinations will undoubtedly bring this further, perhaps to the point where it is clear who the main actors are and how the early thalamus becomes ready to grow to its full size and capabilities.

	Rostral End	Rostral	Caudal	Caudal End
E13.5 area (μm^2)	246000	373000	473000	164000
Adult nuclei count	6	12	9	5
E13.5 protonuclei size (μm^2)	41000	31083	52555	32800
Previous size if square (μm)	202	176	229	181

Table 4.1: Estimations of protonuclei size at E13.5.
First row: measured area of the E13.5 thalamus.
Second: number of nuclei present in the associated representative adult section.
Third: estimations of protonuclei size based on the previous two values assuming size is the same for all.
Fourth: corresponding protonucleus size assuming square shape, calculating the square root of the value in row three. This is useful when examining the images with the help of the scale bar.

4.2 Future work

More relationships

This project has focused on identifying cell populations that co-express pairs of genes to then use the patterns found to study the development of the thalamus.

The same procedure could be applied to find co-expression relationships between groups of three genes or more, and the resulting patterns compared to the already calculated of the two-gene case. It is likely that, at least with the gene data used for this project, not many triplets would provide meaningful results, but at the same time those who did might clearly identify thalamic areas with key importance for the development at that stage or perhaps subdivide the space even further revealing an already present combinatorial genetic profile of cell clusters.

Genes and age

It would be very interesting to expand the analysis of the thalamus-expressed genes both gene and age wise.

The addition of more significative genes would allow the study of further gene pairs and their co-expression patterns might either fall into a previously shown category or provide information to further subdivide the thalamus. A wider range of genes would make it possible to study the resulting subdivision models based on different grouping of genes: for example, how different combinations of guidance cue generating genes parcellate the space if they do.

Another approach to expand the work in this project would be to follow the patterns of co-expression of the presented genes through different developmental time points. Adding key ages such as E15.5, E18.5 and P4 would allow us to understand whether these genes and the thalamic division derived from their co-expression gain complexity over time or they keep the situation similar and rely on other genes to subdivide further.

Mass application

The most appealing and challenging application of the procedures described in this project would be its massive application to all the available gene expression data for a specific tissue area or whole brain.

Properly automated and applied to the whole ADBA data, it would prove a very useful tool to classify gene interactions via co-expression. For a given age and applied to the thalamus, it would be possible to consider all the gene overlap areas to potentially discover common subdivisions that would shed some light on the the early thalamic patterning. The same procedure could be applied to data from other embryonic stages and the same early partitioning of space plus further increase of complexity could be studied over time.

The creation of an automated workflow should be done in close collaboration with the Allen Institute. It would be essential to use some of their tools to manage anatomy correctly: the alignments and graftings shown here have been done manually due to the lack of reference three-dimensional information regarding tissue boundaries. Interaction with such an institution would also guarantee the enormous computational power required, as well as ensuring that the resulting data is readily available to be shared with the scientific community.

4.3 Conclusions

The problem of 3D reconstruction

During a considerable part of this project the three-dimensional reconstruction of the obtained data was a key aspect to develop.

While initially the ADBA is a remarkably big source of data and the reference datasets are very detailed, the number of *in situ* sections is considerably lower. Subsets that span the thalamus do not contain many images, and the number of sections two genes have in common is much smaller. The result of that is reconstructions with big gaps or highly interpolated ones that lose any meaning. It is because of these reasons the representation of results in a 3D space was left aside.

It would be possible to properly reconstruct the reference dataset, or even the thalamus with its 55 sections, but this data, even in limited access form, is already available at the Allen Institute's website. It would be interesting to integrate the density information with their powerful software and data, especially because they have the tools to overcome the problems that the variations of subject tissue and slicing angle systematically cause.

Gene co-expression tools

A wide set of tools have been developed and described to extract information from biological data that is not readily accessible for quantitative purposes.

Even though the procedures have been used only on data from the ADBA, they will probably work with other data as long as the protocols for generating it are similar: nuclei counterstain to extract density information, and *in situ* hybridisation for the gene expression levels. If slicing angles are carefully observed it should be possible to incorporate one's own data into the body of knowledge generated in this project.

Application to the developing thalamus

The described tools have been then applied to the E13.5 thalamus using a pool of genes representing important aspects of development at the time: tissue patterning, cell identity, generation, adhesion and signaling.

The combinations of the gene pair co-expression patterns have revealed an underlying organisation of the thalamus at that stage. Some divisions were expected to be found since thalamic parcellation starts at that time, but the distribution of areas in which cells express different genes and therefore probably have different identity is very detailed.

Past studies (most notably ref. [21]) had described the patterning of the thalamus based on the expression of influential genes in a combinatorial manner. The work shown here is the first attempt to study thalamic parcellation that, using higher quality data, combines informatics tools to involve

more information and takes into account the levels of gene co-expression for all the possible pairs of considered genes as opposed to using single patterns at a time.

This project also shows the way to create powerful screening tools that would make it possible to calculate gene co-expression relationships in any desired way for the tissue area required. This would no doubt be a valuable addition to the biological data *factories* such as the Allen Institute, as the application of these procedures can enhance their contribution to the scientific community.

Appendix

Abbreviations

Abbr.	Name
AD	Anteriodorsal
AM	Anteromedial
BVM	Basal ventromedial
CL	Centrolateral
CM	Centromedial
CeM	Centre median
IAD	Inferior anteriodorsal
IMD	Inferior mediodorsal
LD	Lateral dorsal
LGN(d)	Lateral geniculate (dorsal) nucleus
LP	Lateral posterior
MD	Mediodorsal
MGc	Medial geniculate complex
MV	Medioventral (reuniens)
PC(N)	Paracentral
PF	Parafascicular

Abbr.	Name
PO(m)	Posteromedial
PP	Peripeduncular
PT	Parataenial
Pu	Pulvinar
PVT	Paraventricular
R	Reticular
Rh	Rhomboid
SPF	Subparafascicular
VA	Ventral anterior
VAL	Ventral anterior lateral
VLc	Ventrolateral complex
VM	Ventromedial
VPi	Ventroposterior inferior
VPL	Ventroposterior lateral
VPM	Ventroposterior medial

Table 4.2: Abbreviations used for thalamic nuclei and subdivisions

Relationship to the Allen brain atlas data

Local #	Navigator #	Atlas ID
1	187	199
2	186	198
3	185	197
4	184	196
5	183	195
6	182	194
7	181	193
8	180	192
9	179	191
10	178	190
11	177	189
12	176	188
13	175	187
14	174	186
15	173	185
16	172	184
17	171	183
18	170	182
19	169	181
20	168	180
21	167	179
22	166	178
23	165	177
24	164	176
25	163	175
26	162	174
27	161	173
28	160	172

Local #	Navigator #	Atlas ID
29	159	171
30	158	170
31	157	169
32	-	N/A
33	156	167
34	155	166
35	154	165
36	153	164
37	152	163
38	151	162
39	150	161
40	149	160
41	148	159
42	147	158
43	146	157
44	145	156
45	144	155
46	143	154
47	142	153
48	141	152
49	140	151
50	-	N/A
51	139	149
52	138	148
53	137	147
54	136	146
55	135	145

Table 4.3: Link between local thalamic atlas and the original data from the ADMBA.
The first column indicates section number in the local dataset focused on the thalamus that has been used as reference on which experimental data has been incorporated. The second column is the number the data navigator tool in the ADMBA website associates with that section. The third column is the code that each section is assigned based on the sequential slicing of tissue. Two sections, local 32 and 50, are missing from the full dataset. The navigator does not take them into account, but the local thalamic atlas does.

Publication

Bibliography

[1] A. Agmon and B.W. Connors. Thalamocortical responses of mouse somatosensory (barrel) cortex in vitro. *Neuroscience*, 41(2-3):365–379, 1991.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 2008.

[3] C. Auladell, P. Perez-Sust, H. Super, and E. Soriano. The early development of thalamocortical and corticothalamic projections in the mouse. *Anat Embryol (Berl)*, 201(3):169–179, Mar 2000.

[4] Nataliya Borodovsky, Tatyana Ponomaryov, Shani Frenkel, and Gil Levkowitz. Neural protein olig2 acts upstream of the transcriptional regulator sim1 to specify diencephalic dopaminergic neurons. *Dev Dyn*, 238(4):826–834, Apr 2009.

[5] M. Brown, R. Keynes, and A. Lumsden. *The Developing Brain*. Oxford University Press, 2001.

[6] A. Bulfone, L. Puelles, M.H. Porteus, M.A. Frohman, G.R. Martin, and J.L. Rubenstein. Spatially restricted expression of Dlx-1, Dlx-2 (Tes-1), Gbx-2, and Wnt-3 in the embryonic day 12.5 mouse forebrain defines potential transverse and longitudinal segmental boundaries. *J Neurosci*, 13(7):3155–3172, Jul 1993.

[7] I. Bureau, F. von Saint Paul, and K. Svoboda. Interdigitated paralemniscal and lemniscal pathways in the mouse barrel cortex. *PLoS Biol*, 4(12):e382, Nov 2006.

[8] Y.D. Van der Werf, M.P. Witter, and H.J. Groenewegen. The intralaminar and midline nuclei of the thalamus. anatomical and functional evidence for participation in processes of arousal and awareness. *Brain Research Reviews*, 39:107–140, September 2002.

[9] A. Dufour, J. Seibt, L. Passante, V. Depaepe, T. Ciossek, J. Frisen, K. Kullander, J.G. Flanagan, F. Polleux, and P. Vanderhaeghen. Area specificity and topography of thalamocortical projections are controlled by ephrin/Eph genes. *Neuron*, 39(3):453–465, July 2003.

[10] M. Fabri and H. Burton. Topography of connections between primary somatosensory cortex and posterior complex in rat: a multiple fluorescent tracer study. *Brain Res*, 538(2):351–357, Jan 1991.

[11] P. P. Gao, Y. Yue, J. H. Zhang, D. P. Cerretti, P. Levitt, and R. Zhou. Regulation of thalamic neurite outgrowth by the eph ligand ephrin-a5: implications in the development of thalamocortical projections. *Proc Natl Acad Sci U S A*, 95(9):5329–5334, Apr 1998.

[12] S.F. Gilbert. *Developmental Biology*. Sinauer, 2 edition, December 2003.

[13] G. Gradwohl, C. Fode, and F. Guillemot. Restricted expression of a novel murine atonal-related bhlh protein in undifferentiated neural precursors. *Dev Biol*, 180(1):227–241, Nov 1996.

[14] M. Haidacher. Multiscale nodule detection in CT data. 2005.

[15] M. Ito. Response properties and topography of vibrissa-sensitive VPM neurons in the rat. *J Neurophysiol*, 60(4):1181–1197, Oct 1988.

[16] E.G. Jones. *The Thalamus*. Cambridge University Press, 2nd edition, March 2007.

[17] A. Kataoka and T. Shimogori. Fgf8 controls regional identity in the developing thalamus. *Development*, 135(17):2873–2881, Sep 2008.

[18] K. Kitamura, H. Miura, M. Yanazawa, T. Miyashita, and K. Kato. Expression patterns of Brx1 (Rieg gene), Sonic hedgehog, Nkx2.2, Dlx1 and Arx during *zona limitans intrathalamica* and embryonic ventral lateral geniculate nuclear formation. *Mech Dev*, 67(1):83–96, Sep 1997.

[19] Y. Lim and J.A. Golden. Patterning the developing diencephalon. *Brain Research Reviews*, 53(1):17–26, January 2007.

[20] E. M. Miyashita-Lin, R. Hevner, K. M. Wassarman, S. Martinez, and J. L. Rubenstein. Early neocortical regionalization in the absence of thalamic innervation. *Science*, 285(5429):906–909, Aug 1999.

[21] Y. Nakagawa and D.D. O'Leary. Combinatorial expression patterns of LIM-homeodomain and other regulatory genes parcellate developing thalamus. *J Neurosci*, 21(8):2711–2725, Apr 2001.

[22] X Oliver-Duocastella. Automatic cell counting in 3D *Drosophila* brain scans. Master's thesis, Neuroinformatics Doctoral Training Centre, University of Edinburgh, 2007.

[23] E. Puelles, D. Acampora, R. Gogoi, F. Tuorto, A. Papalia, F. Guillemot, S. Ang, and A. Simeone. Otx2 controls identity and fate of glutamatergic progenitors of the thalamus by repressing GABAergic differentiation. *J Neurosci*, 26(22):5955–5964, May 2006.

[24] Y. Sato, S. Nakajima, N. Shiraga, H. Atsumi, S. Yoshida, T. Koller, G. Gerig, and R. Kikinis. Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images. *Med Image Anal*, 2(2):143–168, Jun 1998.

[25] S. Scholpp, A. Delogu, J. Gilthorpe, D. Peukert, S. Schindler, and A. Lumsden. Her6 regulates the neurogenetic gradient and neuronal identity in the thalamus. *Proc Natl Acad Sci U S A*, 106(47):19895–19900, Nov 2009.

[26] S. Scholpp, I. Foucher, N. Staudt, D. Peukert, A. Lumsden, and C. Houart. Otx1l, Otx2 and Irx1b establish and position the ZLI in the diencephalon. *Development*, 134(17):3167–3176, Sep 2007.

[27] S. Scholpp and A. Lumsden. Building a bridal chamber: development of the thalamus. *Trends Neurosci*, 33(8):373–380, Aug 2010.

[28] J. Seibt, C. Schuurmans, G. Gradwhol, C. Dehay, P. Vanderhaeghen, F. Guillemot, and F. Polleux. Neurogenin2 specifies the connectivity of thalamic neurons by controlling axon responsiveness to intermediate target cues. *Neuron*, 39(3):439–452, Jul 2003.

[29] S. Murray Sherman. The thalamus is more than just a relay. *Curr Opin Neurobiol*, 17(4):417–422, Aug 2007.

[30] S. Murray Sherman and R. W. Guillery. *Exploring the Thalamus and Its Role in Cortical Function: Second Edition*. The MIT Press, 2 edition, December 2005.

[31] L. Sommer, Q. Ma, and D. J. Anderson. neurogenins, a novel family of atonal-related bhlh transcription factors, are putative mammalian neuronal determination genes that reveal progenitor cell heterogeneity in the developing cns and pns. *Mol Cell Neurosci*, 8(4):221–241, 1996.

[32] Y. Suda, Z. M. Hossain, C. Kobayashi, O. Hatano, M. Yoshida, I. Matsuo, and S. Aizawa. Emx2 directs the development of diencephalon in cooperation with otx2. *Development*, 128(13):2433–2450, Jul 2001.

[33] S. C. Suzuki, T. Inoue, Y. Kimura, T. Tanaka, and M. Takeichi. Neuronal circuits are subdivided by differential expression of type-ii classic cadherins in postnatal mouse brains. *Mol Cell Neurosci*, 9(5-6):433–447, 1997.

[34] A. Suzuki-Hirano, M. Ogawa, A. Kataoka, A.C. Yoshida, D. Itoh, M. Ueno, S. Blackshaw, and T. Shimogori. Dynamic spatiotemporal gene expression in embryonic mouse thalamus. *J Comp Neurol*, 519(3):528–543, Feb 2011.

[35] N. Szabó, T. Zhao, X. Zhou, and G. Alvarez-Bolado. The role of Sonic hedgehog of neural origin in thalamic differentiation in the mouse. *J Neurosci*, 29(8):2453–2466, Feb 2009.

[36] H. Takebayashi, S. Yoshida, M. Sugimori, H. Kosako, R. Kominami, M. Nakafuku, and Y. Nabeshima. Dynamic expression of basic helix-loop-helix olig family members: implication of olig2 in neuron and oligodendrocyte differentiation and identification of a new member, olig3. *Mech Dev*, 99(1-2):143–148, Dec 2000.

[37] P. Vanderhaeghen and F. Polleux. Developmental mechanisms patterning thalamocortical projections: intrinsic, extrinsic and in between. *Trends Neurosci*, 27(7):384–391, Jul 2004.

[38] T.Y. Vue, J. Aaker, A. Taniguchi, C. Kazemzadeh, J.M. Skidmore, D.M. Martin, J.F. Martin, M. Treier, and Y. Nakagawa. Characterization of progenitor domains in the developing mouse thalamus. *J Comp Neurol*, 505(1):73–91, Nov 2007.

[39] T.Y. Vue, K. Bluske, A. Alishahi, L.L. Yang, N. Koyano-Nakagawa, B. Novitch, and Y. Nakagawa. Sonic hedgehog signaling controls thalamic progenitor identity and nuclei specification in mice. *J Neurosci*, 29(14):4484–4497, Apr 2009.

[40] K. Yuge, A. Kataoka, A.C. Yoshida, D. Itoh, M. Aggarwal, S. Mori, S. Blackshaw, and T. Shimogori. Region-specific gene expression in early postnatal mouse thalamus. *J Comp Neurol*, 519(3):544–561, Feb 2011.